Analysis and Visualisation of the Metastable States and Pathways Sampled in Molecular Dynamics Simulations

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Nicolas Blöchliger
aus
Unterägeri ZG

Promotionskomitee

Prof. Dr. Amedeo Caflisch (Vorsitz)
Prof. Dr. Madhavi Krishnan
Prof. Dr. Reinhard Furrer
Dr. Andreas Vitalis

Zürich 2015

Summary

In computational biophysics, molecules of biological importance are observed with nearly unlimited spatial and temporal resolutions. The resultant trajectories are often very large data sets of high dimensionality, and this demands dedicated and scalable algorithms for their analysis and visualisation. These analyses are typically expected to identify metastable conformations of biomolecules and to describe the connectivities among these states. In practice, the initial mining of simulation data is often exploratory in nature and can draw from a wide range of algorithms developed in the data and information sciences. However, many of these algorithms do not scale to large data sets or risk the loss of important information.

As a cornerstone of this thesis, we present a novel analysis tool to mine data from molecular simulations. The main output of this method is the Sapphire (States And Pathways Projected with HIgh REsolution) plot, which reveals metastable states and sequences of events sampled in these simulations. Specifically, the individual observations or snapshots are first arranged in a way that groups them in correspondence with distinct regions of high sampling density. The resultant sequence is annotated to highlight, distinguish, and characterise these different regions, which are interpreted to be free energy basins.

The Sapphire plot is an excellent visual tool for exploratory data analysis and communication of results in a single figure. Thermodynamic information about proteins, which are systems of considerable complexity, is resolved quantitatively. Given a sufficient overall sampling density, distinct states do not overlap, and every snapshot is represented, i.e., resolution is maximal. In addition, the pathways and events sampled by simulations can be visualised efficiently. The Sapphire plot can be computed in $\mathcal{O}(N \log N)$ time with an approximate algorithm, where N is the number of snapshots. This means that the method is scalable to very large data sets that can be produced quickly and in routine fashion on present-day supercomputers.

We have used Sapphire plots to analyse data from molecular dynamics simulations of protein folding, of the conformational dynamics within folded states of proteins, and of the binding of a peptide to a folded protein domain. Particular emphasis is given to the last application. Here, the Sapphire plot and subsequent analyses revealed the mechanism of the binding process in exceptional detail. In particular, we observed the fast formation of an encounter complex driven by electrostatic interactions. The Sapphire plot showed that this complex is disordered and stabilised by non-native interactions, and

that its formation precedes the rate-limiting step of binding. In addition, our analysis tools allowed us to identify multiple binding modes within the specific complex.

SAPPHIRE plots - as well as most other data analysis protocols - depend on how the input data are represented, and on how distance between snapshots is measured. If the raw data are of high dimensionality, unexpected problems can arise. Specifically, a large number of irrelevant dimensions or features can mask information related to processes of interest. The simple exclusion of features will be ineffective if the importance of different features varies among the data points. As a consequence, finding suitable distance functions "by hand" can be a time-consuming task, and the choice of distance function often has a greater impact than the choice of the method used to mine the data.

To address these difficulties, we introduce the weighting of features based on their autocorrelation function or on time-local measures of memory loss. Our tests demonstrate that additional information can be extracted from molecular dynamics trajectories if degrees of freedom such as side chains, flexible loops, and terminal residues are included with appropriate weights. These results challenge the existing paradigm of creating "good" distance functions by discarding features deemed to be too noisy or uninteresting.

Zusammenfassung

Die computergestützte Biophysik nutzt die Technik der Moleküldynamiksimulationen, um Biomoleküle mit nahezu unbeschränkter räumlicher und zeitlicher Auflösung zu beobachten, und dies erzeugt riesige Datenmengen. Solche Trajektorien sind sowohl (zeitlich) lang als auch (räumlich) hochdimensional und ihre Analyse erfordert daher massgeschneiderte und skalierbare Algorithmen. Das Ziel solcher Analysen ist wesentlich die Identifikation der metastabilen Zustände des simulierten Systems und die Beschreibung ihrer Vernetztheit. Die Datenwissenschaft und verwandte Disziplinen stellen bereits eine Vielzahl von Methoden bereit. Erhebliche Einschränkungen dieser Methoden liegen im Verlust wichtiger Informationen und einer oft ungenügenden Skalierbarkeit.

Der Kern dieser Dissertation ist die Präsentation einer neuen Methode zur Analyse von aus Moleküldynamiksimulationen gewonnenen Daten. Das SAPPHIRE-Diagramm (States And Pathways Projected with HIgh REsolution) ist dabei das eigentliche Hauptergebnis unserer Methode. Es zeigt die metastabilen Zustände sowie die zeitlichen Prozessabläufe des simulierten Systems detailliert auf. Um das Diagramm zu generieren werden die Datenpunkte so angeordnet, dass diejenigen, die demselben Häufungspunkt oder Dichtemaximum im Datensatz entsprechen, nahe aufeinanderfolgen. Die resultierende Folge von Datenpunkten wird daraufhin mit zusätzlichen Informationen versehen, die verschiedene Häufungspunkte auf vielfältige Art und Weise unterscheiden und charakterisieren. Bei geeigneter Repräsentation der Daten sollten die einzelnen Häufungspunkte den metastabilen Zuständen des Systems entsprechen.

Unsere Tests haben ergeben, dass SAPPHIRE-Diagramme für die explorative Analyse von Moleküldynamiksimulationen äusserst nützlich sind. Unterschiedliche metastabile Zustände überlagern sich nicht, solange die allgemeine Datendichte hoch genug ist. Thermodynamische und kinetische Informationen können effizient visualisiert werden. Da jeder einzelne Datenpunkt im Diagramm wiedergegeben wird, ist die Auflösung maximal. Wir haben einen approximativen Algorithmus zur Berechnung von SAPPHIRE-Diagrammen entwickelt, der eine Laufzeit von $\mathcal{O}(N \log N)$ aufweist. Grosse Datenmengen, die mit modernen Supercomputern schnell und kostengünstig generiert werden können, sind daher mit diesem Algorithmus problemlos zu analysieren.

Wir haben SAPPHIRE-Diagramme verwendet um verschiedene biomolekulare Prozesse wie die Proteinfaltung oder andere Formen von Strukturfluktuationen zu untersuchen. Ein Schwerpunkt lag dabei in der Frage wie Peptide an Proteine binden. Für unser Fallbeispiel beobachteten wir eine schnelle Assoziation der beiden Moleküle aufgrund von

elektrostatischer Anziehung. Das SAPPHIRE-Diagramm zeigte auf, dass der daraus hervorgehende Komplex ungeordnet ist und durch Salzbrücken, die in der finalen Anordnung nicht vorkommen, stabilisiert wird. Ein weiteres Resultat war die Unterscheidung von Substrukturen des stereospezifischen Komplexes.

Wie bei den meisten Analysemethoden hängt auch bei SAPPHIRE-Diagrammen das Endergebnis davon ab, wie die Eingabedaten repräsentiert werden oder wie die Distanz zwischen Datenpunkten definiert ist. Hochdimensionale Daten führen dabei zu besonderen Komplikationen. So kann die Präsenz zu vieler unerheblicher Variablen Informationen über die im Vordergrund der Analyse stehenden Prozesse verbergen. Besonders wenn sich die Relevanz von Variablen zwischen Datenpunkten unterscheidet, greift eine simple Einteilung in relevante und in unerhebliche Variablen allerdings zu kurz. Als Folge dessen kann die manuelle Konstruktion einer angemessenen Distanzfunktion zeitraubend und unergiebig sein. Das ist von erheblicher Bedeutung, da die Wahl der Distanzfunktion oft entscheidender ist als die Wahl der Analysemethode selbst.

In der vorliegenden Arbeit präsentieren wir eine effiziente Methode zur globalen oder zeitlich-lokalen Gewichtung von Variablen, die viele der obengenannten Probleme umgeht. Unsere Untersuchungen haben ergeben, dass mit Hilfe gewichteter Distanzfunktionen zusätzliche Informationen aus den Simulationsdaten gewonnen werden können. Diese Resultate stellen die gängige Praxis in Frage, die durch das blosse Ignorieren von vermeintlich unerheblichen Variablen informative Distanzfunktionen zu konstruieren versucht.

Contents

Chapter 1

Introduction

The ubiquitous, large and high-dimensional data sets encountered in many scientific studies [1–7] demand suitable, scalable algorithms for exploratory analysis [8–11], visualisation [11–15], and unsupervised learning [2, 16–19]. In computational biophysics, atomistic simulations are routinely used to generate long and high-dimensional trajectories of biomolecules [5, 6, 18, 20, 21], and we present here a novel analysis tool [22, 23] called Sapphire plot to mine such data sets.

Our main interest lies in simulations of the dynamics of proteins, and we subsequently review their properties as well as the methodology of molecular dynamics (MD) simulation. We then survey the challenges of interpreting molecular simulation data and describe Sapphire plots, which are comprehensively examined, tested, and applied in Chapters 2–5. The outcome of most data analysis algorithms depends on how the input data are represented, especially if they are high-dimensional. We touch upon methods for data preprocessing in the context of biomolecular simulation towards the end of this chapter. An extensive discussion is provided in Chapter 5.

1.1 Proteins

Proteins are biological macromolecules involved in a huge variety of biochemical processes. For example, enzymes catalyse chemical reactions, receptors participate in signalling, transcription factors regulate gene expression, ion channels control transport across the cell membrane, and muscle proteins enable movement. Studying proteins is thus essential to understanding life at the molecular level.

1.1.1 Structure

Proteins are composed of one or more polypeptide chains. The building blocks of these polymers are amino acid residues, which consist of a central carbon atom, termed C_α atom, linked to an amino group (NH_2), a hydrogen atom, a side chain, and a carboxyl group (COOH). Biological polypeptides range in size from a few amino acids to several

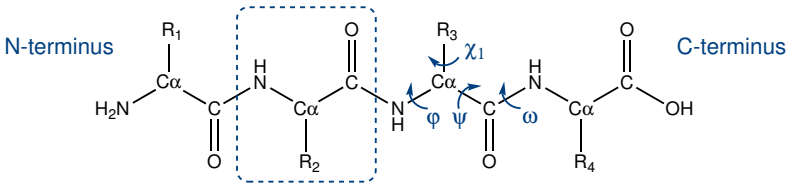


Figure 1.1: **Structure of a peptide.** The peptide consists of four amino acid residues, the second of which is highlighted. The residues are linked by peptide bonds (C–N), and R_1, R_2, R_3 and R_4 represent their side chains. Arrows indicate the four types of dihedral angles found in peptides and proteins.

thousands [24]. Extremely large proteins more than 1 μm in length and composed of up to 30000 amino acids have been identified [25].

Amino acids are connected in linear fashion by peptide bonds linking the carboxyl group of one amino acid to the amino group of the next. The two ends of a polypeptide chain are called N-terminus and C-terminus (Fig. 1.1). Rotation around peptide bonds is hindered by electronic effects. The overall conformation of the backbone of a protein, i.e., the protein without its side chains, can thus be described by the φ and ψ dihedral angles of each amino acid [24]. Dihedral angles describe torsion around a covalent bond. In the context of proteins, there are four different types of dihedral angles (Fig. 1.1). The φ, ψ, and ω angles describe rotation around the N–C_α bond, the C_α–C bond, and the peptide bond (C–N), respectively. The $\chi_1, \ldots, \chi_n$ angles describe the rotameric states of the side chains, where n depends on residue type.

Under physiological conditions, many proteins adopt a well-defined structure referred to as the native or folded conformation. It is encoded in the sequence of the protein's constituting amino acids and their chemical properties (e.g., polarity, charge, and size) [26], and it can be determined experimentally by nuclear magnetic resonance spectroscopy [27] or X-ray crystallography [28–30]. For soluble proteins, polar and charged side chains tend to be on the surface of the protein while nonpolar side chains are mostly buried and form what is called the hydrophobic core. Inspection of folded structures reveals recurring patterns in the conformation of the protein backbone, the most prevalent of which are the α-helix and the β-sheet (Fig. 1.2) [24]. The α-helix denotes a helical conformation with characteristic backbone hydrogen bonds between residue i and residue $i + 4$ and side chains pointing away from the helix axis and towards its N-terminus. The α-helix is thus a local structural element in the sense that it involves backbone interactions between residues close in sequence. The β-sheet on the other hand involves hydrogen bonds between residues that can be far in sequence or even belong to distinct polypeptides. Here, hydrogen bonds are formed between the polar backbone groups of extended segments of the polypeptide called β-strands. Neighbouring β-strands can be aligned either in parallel or antiparallel fashion. Parallel β-sheets are generally less stable due to unfavourable distortion of the hydrogen bonds. These backbone arrangements are referred to as the protein's secondary structure, and the arrangement of the different secondary structure elements with respect to each other is termed its tertiary structure.

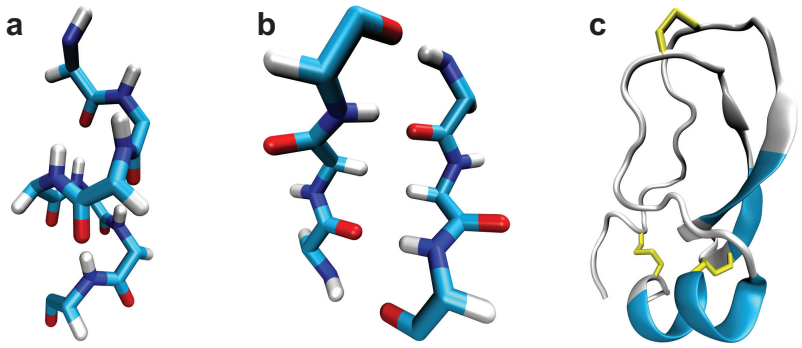


Figure 1.2: **Protein secondary and tertiary structure.** Secondary structure elements like α-helices **(a)** and β-sheets **(b)** exhibit characteristic backbone hydrogen bonds. Carbon, nitrogen, oxygen, and hydrogen atoms are shown in cyan, blue, red, and white, respectively. For clarity, side chains are not shown. **(c)** Tertiary structure of bovine pancreatic trypsin inhibitor (BPTI). The cartoon illustration traces the backbone of the protein, and secondary structure elements are shown as ribbons. The residues shown in (a) and (b) are indicated in cyan. Cysteine side chains are shown in yellow, and their hydrogen atoms are omitted for clarity.

Pairs of cysteine residues can be linked by covalent disulfide bonds through oxidation of their thiol groups (SH) [31]. The presence of disulfide bonds imposes constraints on the protein structure and can have various consequences, including stabilisation of the folded conformation. For example, the protein bovine pancreatic trypsin inhibitor (BPTI), studied below and in Chapters 3 and 5, has three disulfide bonds (Fig. 1.2c) the isomerisations of which are coupled to conformational dynamics in the native state [32].

1.1.2 Dynamics

Proteins are flexible molecules at physiological temperatures, and their dynamics cover a wide range of time and length scales [33]. High-frequency bond vibrations (on a time scale of about 10^{-14} s) and methyl group rotations (10^{-11} s) are orders of magnitude faster than rotation of surface side chains (10^{-9} s) and loop motion (10^{-8} s). Rotation of buried side chains (10^{-5} s) and large-scale transitions like domain motion or complete unfolding and refolding are again orders of magnitude slower ($> 10^{-5}$ s). There is of course considerable heterogeneity in these time scales among different proteins. Protein folding rates, for example, span eight orders of magnitude [34]. In the context of the work presented here, we are interested in three processes involving proteins, namely protein folding, conformational dynamics in the native state, and protein-protein binding (Fig. 1.3).

First, protein folding denotes the process by which an extended polypeptide chain adopts its native conformation. In vivo this occurs during or after synthesis by the ribosome. In vitro refolding happens after chemical or heat-induced unfolding. The protein folding problem has several distinct aspects, e.g., prediction of the native structure of a

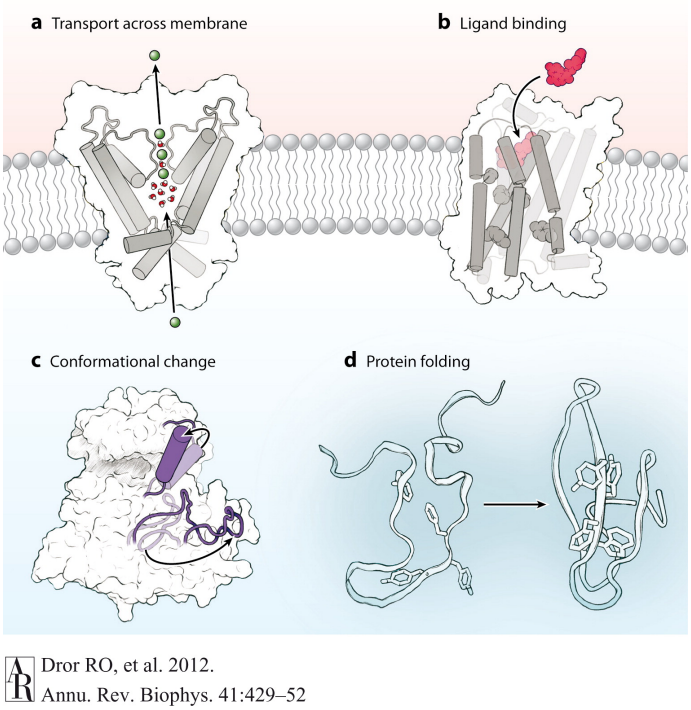


Figure 1.3: **Proteins are involved in a number of biologically relevant processes.** These include **(a)** transport across the cell membrane, **(b)** ligand binding, **(c)** conformational dynamics in the folded state, and **(d)** protein folding. See main text for a discussion of the processes depicted in (b), (c) and (d). Reprinted from Ref. [21].

protein from its amino acid sequence [35,36] and Levinthal's paradox [37,38], which asks how a protein manages to find its native structure on biologically relevant time scales despite the size of the available conformational space. While some researchers have begun to argue that the protein folding problem is solved in terms of general principles, it is clear that numerous questions remain [34,39–42]. We note that the reverse process, protein unfolding, is biologically relevant as well given that partial unfolding has been linked to signalling [43] and assisted unfolding plays a role in protein degradation [44].

Second, at finite temperatures, the native conformation of a protein is not fully rigid. Conformational dynamics and changes therein are essential for processes related to protein function [33, 45], e.g., catalysis [46–49], signalling [50], and allostery [51]. Characterising protein dynamics and their link with function remains an active area of research.

Third, biomolecules give rise to living matter by interacting with each other. Specific protein-protein binding, one instance of biomolecular interactions, can be described as diffusional association followed by formation of a stereospecific complex [52, 53]. Open challenges include characterisation of the encounter complex [54] and answering the question how proteins manage to bind their targets both rapidly and specifically in a crowded

biological cell [55,56].

It should be noted that these three processes are all driven by weak non-bonded interactions [57], and that they can be coupled. For example, protein dynamics are important for binding [58–60] and disordered proteins that only fold upon binding their target have been identified [61].

1.1.3 The physics of proteins

We next describe proteins from a physicist's point of view. After recapitulating the basic concepts from thermodynamics and kinetic theory, we summarise mechanistic models for protein folding, the concept of diffusion on the free energy surface, and relevant aspects of the theory of stochastic processes.

Thermodynamics Biophysical processes at constant pressure and temperature are governed by the Gibbs free energy $G = H - TS$ [24]. For a biochemical process at constant pressure, the change in enthalpy ΔH is equal to the heat absorbed or released and ΔS is the change in entropy, which is a measure of relative disorder. Spontaneous reactions require $\Delta G = \Delta H - T\Delta S \leq 0$ according to the laws of thermodynamics. At equilibrium, $\Delta G = 0$ and $\Delta G^{\ominus} = -RT \ln K$, where $\Delta G^{\ominus}$ is the free energy change of the reaction when its reactants and products are in their standard states, K is the equilibrium constant, and $R = 8.31\ \mathrm{J/mol\,K}$ is the gas constant.

It is important to note that contributions of the solvent are crucial. For example, folding of a globular protein is driven by the hydrophobic effect and hydrogen bonding, while the loss of conformational entropy of the protein destabilises the folded state [62]. The hydrophobic effect is the sum of favourable van der Waals interactions due to tight packing of hydrophobic side chains in the protein core and an increase in solvent entropy. Since an unfolded protein can form hydrogen bonds with water, it is not completely clear to what extent protein-protein hydrogen bonds stabilise the folded state [62,63]. Under physiological conditions, these opposing tendencies result in a conformational stability (with respect to the denaturation midpoint) of the folded state that is typically around -5 to -15 kcal/mol [64].

Kinetics The simplest description of protein dynamics is given by the two-state mechanism

$$A \rightleftharpoons B,$$

where both the forward and reverse reaction are first order with rate constants k_{AB} and k_{BA}. Here, A and B could for example denote the folded and unfolded states of a protein. The probability P_A to find the system in state A and the probability P_B for state B evolve according to

$$\begin{aligned} \frac{dP_A}{dt} &= k_{BA}P_B(t) - k_{AB}P_A(t), \\ \frac{dP_B}{dt} &= k_{AB}P_A(t) - k_{BA}P_B(t). \end{aligned} \tag{1.1}$$

At equilibrium, $\frac{dP_A}{dt} = \frac{dP_B}{dt} = 0$, and therefore $P_A^{eq} = k_{BA}/(k_{AB} + k_{BA})$ and $P_B^{eq} = k_{AB}/(k_{AB} + k_{BA})$. Thus, the equilibrium constant satisfies $K = P_A^{eq}/P_B^{eq} = k_{BA}/k_{AB}$.

The time-dependent solution of Eq. 1.1 is given by

$$P_A(t) = P_A^{eq} + (P_A(0) - P_A^{eq})e^{-kt}, \tag{1.2}$$

where $k = k_{AB} + k_{BA}$. Lifetimes in the two states follow the exponential distributions $p(t_A) = k_{AB}e^{-k_{AB}t_A}$ and $p(t_B) = k_{BA}e^{-k_{BA}t_B}$ with mean lifetimes given by $\tau_A = k_{AB}^{-1}$ and $\tau_B = k_{BA}^{-1}$.

The two-state model is only appropriate if there is a separation of time scales, i.e., if the dynamics within the two states occur on time scales much shorter than τ_A and τ_B, and if transitions between A and B can be assumed to be instantaneous. It is of course possible to include (on-pathway as well as off-pathway) intermediate states to model transitions between A and B, but the same caveats apply to multi-state descriptions.

The autocorrelation of any ideal noise-free observable with distinct values in states A and B is given by e^{-kt}. This explains the name "observed rate" for k. Experimental data is often interpreted in terms of such exponential relaxation. While theoretical arguments for two-state folding have been advocated [65], the usual argument for such behaviour is simply that a two-state model is sufficient to explain the data. However, computational studies have pointed out multiple pathways, additional metastable conformations, and unfolded states with residual structure and slow reconfiguration times [66–69]. More recently, experimental evidence for unexpected complexity in the native state, multiple folding pathways, and kinetic networks of multiple metastable conformations has been found as well [70–73].

Mechanistic models for protein folding Several mechanistic models have been proposed to describe protein folding [74]. In the hydrophobic-collapse model, the protein is thought to first collapse rapidly and to bury hydrophobic side chains nonspecifically before secondary structure elements are formed [75]. Such a collapse leads to a heterogeneous state stabilised by nonnative interactions. The protein is then thought to rearrange its core while remaining in a globular conformation until the native state is found. The diffusion-collision model, on the other hand, predicates that a protein consists of microdomains that can rapidly and randomly sample their available conformations [76,77]. Protein folding is thus reduced to diffusional encounter and coalescence of microdomains, potentially via multiple pathways through metastable intermediates. According to the nucleation model, one or more local, native-like structures form first and subsequently serve as nuclei for propagation of native structure [78]. New mechanistic models have been introduced and combined with older ones as more and more experimental data became available, e.g., the nucleation-condensation model [79], the topomer search model [80], and the zipping-and-assembly model [81]. As of today, mechanistic models have not lost their appeal [82], but none of them is sufficient to account for and predict folding routes and rates of different proteins in general [42].

The free energy surface Considering protein dynamics as diffusion on the free energy surface has served to motivate several important concepts [83]. Here, the free energy G is taken as a function of the configuration of the system and the experimental conditions.

First, it has been recognised that the free energy surface of a protein is hierarchically organised, with different "valleys" exhibiting minima within minima [84]. Second, attention has been given to the roughness of the free energy surface [85] and how it affects kinetics [86]. Third, polypeptide chains with random amino acid sequences have been found to be complex systems in a state of frustration. This means that, like spin glasses, random polypeptides are characterised by multiple low-energy conformations with transitions among them being so slow at low enough temperatures that the system is essentially frozen [85]. Fourth, the concept of a funnel-shaped free energy surface has been employed to describe how proteins resolve Levinthal's paradox and how enthalpy and entropy are balanced during folding [87–89]. Natural proteins are minimally frustrated according to this picture [90], and their free energy surface is shaped in such a way that diffusion on it naturally leads to the native state [66,85,91,92]. The protein thus avoids searching the vast majority of the high-dimensional conformational space. Even very simple models show that a bias towards locally native conformations is sufficient to fold a protein on biological time scales [38]. Recently, experiments have uncovered frustration in natural protein domains [93], and the relation between frustration and function has become a topic of interest [94].

Stochastic processes The theory of stochastic processes gives a quantitative framework for the concepts just discussed. We briefly review some aspects of stochastic processes next [95].

A stochastic process Y is a set of random variables Y_t indexed by time, and we denote the probability density that Y assumes the values $y_1, \ldots, y_n$ at times $t_1 \leq \cdots \leq t_n$ by $P(y_1, t_1; \ldots; y_n, t_n)$. Furthermore, let $P_{l|k}(y_{k+1}, t_{k+1}; \ldots; y_{k+l}, t_{k+l} | y_1, t_1; \ldots; y_k, t_k)$ be the probability density that Y has the values $y_{k+1}, \ldots, y_{k+l}$ at times $t_{k+1} \leq \cdots \leq t_{k+l}$ given that it has the values $y_1, \ldots, y_k$ at times $t_1 \leq \cdots \leq t_k$ where $t_k \leq t_{k+1}$. If

$$P_{1|n-1}(y_n, t_n | y_1, t_1; \ldots; y_{n-1}, t_{n-1}) = P_{1|1}(y_n, t_n | y_{n-1}, t_{n-1}) \tag{1.3}$$

holds for times $t_1 \leq \cdots \leq t_n$, the stochastic process Y is called a Markov process. Equation 1.3 means that the conditional probability for y_n is fully determined by the value y_{n-1} at time t_{n-1} and not affected by the values of Y at previous times. A stochastic process is called stationary if its moments satisfy $\langle Y_{t_1+\tau} \ldots Y_{t_n+\tau} \rangle = \langle Y_{t_1} \ldots Y_{t_n} \rangle$ for all $t_1, \ldots, t_n$ and τ.

For stationary Markov processes, $P_{1|1}(y_2, t_2 | y_1, t_1)$ depends on t_1 and t_2 only through their difference $\tau = t_2 - t_1$, and transition probabilities are defined as

$$T_\tau(y_2 | y_1) = P_{1|1}(y_2, t_2 | y_1, t_1). \tag{1.4}$$

These transition probabilities obey the Chapman-Kolmogorov equation that states that the probability to go from y_1 at time t_1 to y_3 at time t_3 is identical to the probability to

go from y_1 to y_3 via any y_2 at time t_2. For small τ,

$$T_\tau(y_2|y_1) \approx (1 - a(y_1)\tau)\delta(y_2 - y_1) + \tau W(y_2|y_1), \tag{1.5}$$

where $\delta(y_2 - y_1) = 1$ if $y_2 = y_1$ and $\delta(y_2 - y_1) = 0$ else, $1 - a(y_1)\tau$ is the probability that y_1 is not left during τ, and $W(y_2|y_1)$ denotes the transition probability per unit time to go from y_1 to y_2. Based on Eq. 1.5 the master equation

$$\frac{\partial P(y,t)}{\partial t} = \int \left[W(y|y')P(y',t) - W(y'|y)P(y,t)\right] \mathrm{d}y' \tag{1.6}$$

can be derived. The master equation describes the evolution of $P(y,\cdot)$ in terms of gain, $W(y|y')P(y',t)$, and loss, $W(y'|y)P(y,t)$. An example of a master equation for a discrete state space was encountered above (Eq. 1.1).

The Fokker-Planck equation can be derived from the master equation assuming that $W(y|y')$ decays rapidly as a function of $r = y - y'$ (i.e., only small jumps occur), that $W(y|y')$ varies slowly as a function of y', and that $P(y,t)$ varies slowly as a function of y. In one dimension it is given by

$$\frac{\partial P(y,t)}{\partial t} = -\frac{\partial}{\partial y}A_1(y)P(y,t) + \frac{1}{2}\frac{\partial^2}{\partial y^2}A_2(y)P(y,t), \tag{1.7}$$

where $A_i(y) = \int r^i W(y;r)\,\mathrm{d}r$ and $W(y;r) = W(y' = y - r|y)$. The first term on the right-hand side of Eq. 1.7 describes drift and the second one diffusion.

The solution of the master equation or the Fokker-Planck equation can, under a set of assumptions, be written in terms of the eigenfunctions φ_i of the associated operator as

$$P(y,t) = \sum_{i=0}^{\infty} a_i e^{-\lambda_i t}\varphi_i(y). \tag{1.8}$$

Here, the coefficients a_i depend on the initial condition $P(y,0)$, the eigenvalues satisfy $\lambda_0 = 0 < \lambda_1 \leq \lambda_2 \leq \ldots$, and $\varphi_0 = P^{eq}(y) = \lim_{t\to\infty} P(y,t)$ is the distribution at equilibrium. Note that the solution to the two-state problem discussed above (Eq. 1.2) has this form. In the presence of a spectral gap $\lambda_k \ll \lambda_{k+1}$ for some k, the solution P can be approximated as $P(y,t) = \sum_{i=0}^{k} a_i e^{-\lambda_i t}\varphi_i(y)$.

For diffusion of an overdamped particle (high friction limit) in an external potential F, the Fokker-Planck equation reads

$$\frac{\partial P(y,t)}{\partial t} = \frac{\partial}{\partial y}\left(D(y)\left(\frac{1}{k_B T}\frac{\partial F(y)}{\partial y}P(y,t) + \frac{\partial P(y,t)}{\partial y}\right)\right), \tag{1.9}$$

where D is the position dependent diffusion coefficient and $k_B = 1.38 \times 10^{-23}\,\mathrm{J/K}$ is the Boltzmann constant. The Boltzmann distribution, $P^{eq}(y) \propto e^{-F(y)/k_B T}$, is the stationary solution of Eq. 1.9. Equivalently, the dynamics can be described with the

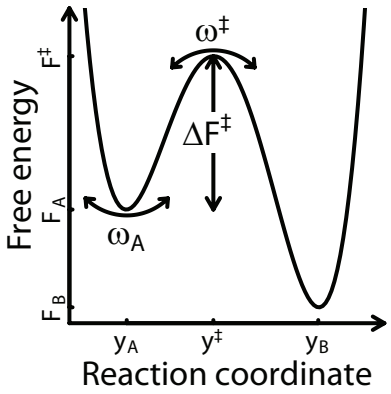


Figure 1.4: **Diffusion in a one-dimensional potential.** Kramers' law (Eq. 1.11) gives the rate of escape from the free energy minimum at y_A over the barrier at $y^\ddagger$ [96]. In the context of protein folding, where Kramers' rate theory has been applied widely, A and B are the unfolded and the folded state and $y^\ddagger$ is the position of the transition state. ω_A and $\omega^\ddagger$ describe the curvature of the free energy profile.

Langevin equation

$$\frac{\partial y(t)}{\partial t} = -\frac{D(y(t))}{k_B T}\frac{\partial F(y(t))}{\partial y} + \sqrt{2D(y(t))}R(t), \tag{1.10}$$

where $R(t)$ is a delta-correlated Gaussian process with zero mean. Kramers used such descriptions to study the escape process from a deep minimum in a one-dimensional potential. For diffusion in a one-dimensional double-well potential with minima at y_A and y_B and a high free energy barrier at $y^\ddagger$ ($\Delta F = F^\ddagger - F_A \gg k_B T$), the rate of escape from the minimum at y_A is according to Kramers' law approximately

$$k = \frac{\omega_A \omega^\ddagger D(y^\ddagger)}{2\pi k_B T} e^{-\Delta F/k_B T}, \tag{1.11}$$

where $\omega_A^2 = \frac{\partial^2 F(y_A)}{\partial y^2}$ and $(\omega^\ddagger)^2 = -\frac{\partial^2 F(y^\ddagger)}{\partial y^2}$ describe the curvature of the free energy profile (Fig. 1.4) [96].

Whether protein dynamics can be described satisfyingly in theory or in practice as a discrete Markov process with a few states or as diffusion in a low-dimensional space remains a matter of debate and a field of ongoing research. For example, experimental and simulation data on protein folding has often been described as diffusion in a one-dimensional potential [5, 93, 97–111]. However, numerous computational and experimental studies have unmasked systems whose complexity cannot be captured with a one-dimensional coordinate [5, 69, 70, 72, 73, 112]. Small proteins have thus been hypothesised to contain at the same time too many degrees of freedom to be characterised with a one-dimensional reaction coordinate (like a reaction of small molecules) and too few to be described by a single order parameter (like a phase transition) [113]. The applicability

of one-dimensional diffusion therefore seems to be the exception rather than the rule.

1.2 Molecular dynamics simulation

The benefits of modelling in biophysics and structural biology are manifold, and computer simulation has established itself as an indispensable complement to theory and experiment [114]. The main motivation for simulating biomolecules is that computational methods offer a temporal and spatial resolution that is difficult to achieve experimentally (Fig. 1.5) [21, 115]. Furthermore, most experimental techniques only provide data averaged over large ensembles while simulations can report on distributions and follow individual molecules. Simulations are thus used to generate structural models or ensembles of structures to interpret experimental data [116], e.g., in the context of X-ray crystallography [117], electron microscopy [118], or nuclear magnetic resonance spectroscopy [119, 120]. On the other hand, simulations are performed to generate experimentally testable hypotheses, and data from simulations of this type are the ones of interest in the present work.

1.2.1 Method

The processes we are interested in can be modelled with reasonable accuracy at the atomic level since they are governed by nonbonded interatomic interactions [121]. Thus, based on the Born-Oppenheimer approximation, the system of interest is represented as a set of atoms and described by their positions, r_i, and velocities, $\dot{r}_i$. The system is propagated according to an appropriate equation of motion such as Newton's second law $F_i = m_i \ddot{r}_i$, where m_i denotes the mass of atom i, $F_i = -\partial U / \partial r_i$ is the force acting on atom i, and U is the potential energy [116, 121–123].

U is modelled with an empirical potential function that most often has the following general form [124, 125].

$$\begin{aligned} U(r) = & \sum_{\text{bonds}} k_b (b - b_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{dihedrals}} k_\chi (1 + \cos(n\chi - \delta)) + \sum_{\text{impropers}} k_{\text{imp}} (\varphi - \varphi_0)^2 \\ & + \sum_{\text{nonbond}} \varepsilon_{ij} \left((r_{ij}^{\min} / r_{ij})^{12} - 2 (r_{ij}^{\min} / r_{ij})^6 \right) + q_i q_j / 4\pi\varepsilon r_{ij} \end{aligned} \tag{1.12}$$

The first four terms in Eq. 1.12 describe bonded interactions. Bond lengths, bond angles, and improper dihedral angles (introduced to describe out-of-plane distortions) are modelled by simple harmonic potentials with force constants k_b, k_θ, and k_{imp} and equilibrium values b_0, θ_0, and φ_0. The term for dihedral angles is based on force constants k_χ, multiplicities n, and phase angles δ. The last term in Eq. 1.12 describes nonbonded interactions as a function of r_{ij}, the distance between two atoms i and j. The Lennard-Jones

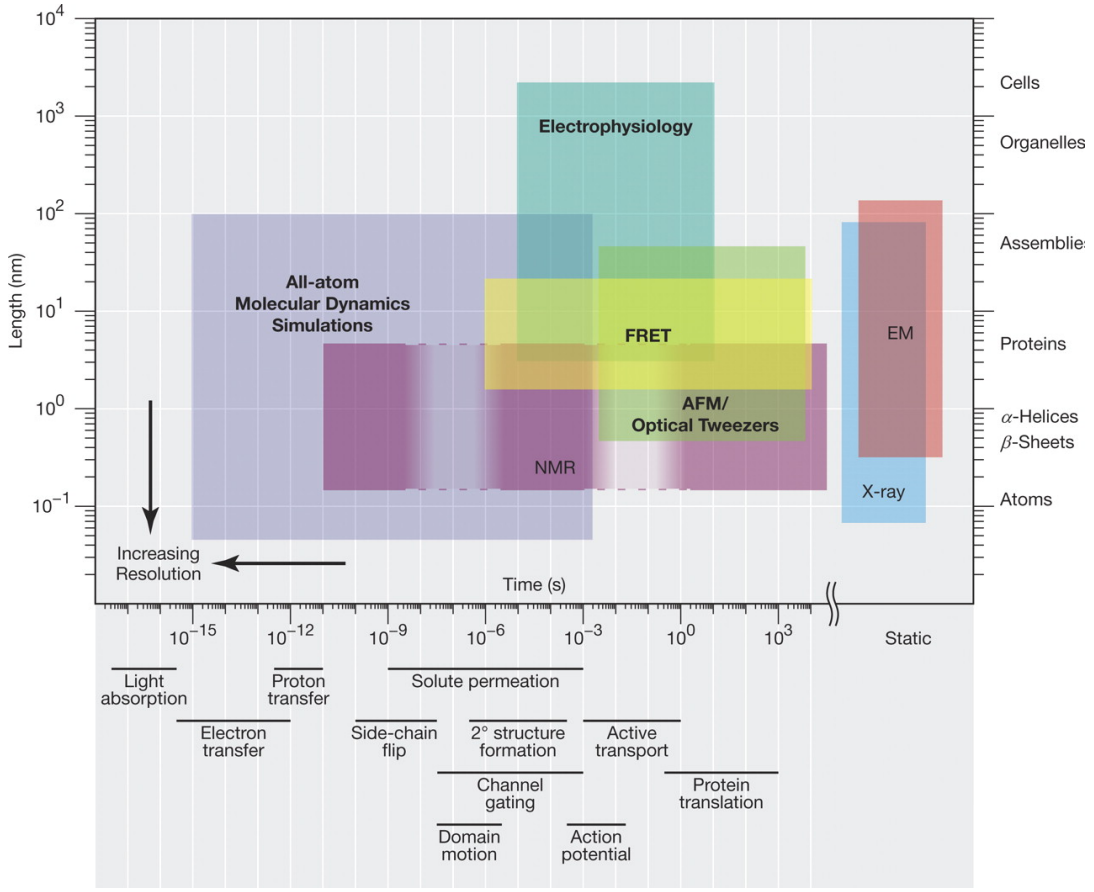


Figure 1.5: **Biophysical techniques vary in their resolution.** Coloured boxes show the spatial and temporal resolution for the individual techniques. The time scales of selected biophysical processes are indicated below the plot. AFM, atomic force microscopy; EM, electron microscopy; FRET, Förster resonance energy transfer; NMR, nuclear magnetic resonance. Reprinted from Ref. [115].

potential

$$\varepsilon_{ij}\left((r_{ij}^{\min}/r_{ij})^{12} - 2(r_{ij}^{\min}/r_{ij})^{6}\right)$$

contains an attractive term accounting for London dispersion forces (r_{ij}^{-6}) and a term modelling repulsion of the electronic shells (r_{ij}^{-12}). The Lennard-Jones potential attains its minimum value $-\varepsilon_{ij}$ for $r_{ij} = r_{ij}^{\min}$. Electrostatic interactions are taken into account by the term $q_i q_j/\varepsilon r_{ij}$, where q_i is the partial charge of atom i and ε is the effective dielectric constant.

The various parameters determining U are optimised to match experimental data and quantum mechanical calculations. Inaccuracies in the potential energy function lead to systematic errors. Current potential energy functions of the form given by Eq. 1.12 are overall sufficiently accurate to predict the folded state of several small proteins, but they

generally fail to reproduce experimental data concerning unfolded conformations [126].

The number of nonbonded interactions in Eq. 1.12 scales quadratically with system size. Truncation schemes are used in conjunction with periodically updated neighbour lists for the short-range interactions described by the Lennard-Jones potential, and scalable algorithms like the particle mesh Ewald summation method [127] are used to calculate the electrostatic interactions [123,128,129]. These techniques entail modifications of Eq. 1.12 and keep the evaluation of U feasible for large systems.

While the first MD simulation of a protein was performed in vacuo [130], it is now standard to solvate the molecules of interest, which requires special treatment of the boundary of the simulation box. In periodic boundary conditions, the simulation box is infinitely but virtually replicated in every direction [123,128,129,131]. The minimum image convention demands that short-range interactions between two atoms i and j are only counted once, namely between i and j', where j' is either j or one of its periodic images, whichever is closest to i.

Numerical solution of the ordinary differential equation $-\partial U/\partial r_i = m_i \ddot{r}_i$ requires initial values for the atomic positions r and velocities $\dot{r}$ as well as a numerical integrator. The starting structure, i.e., the initial value for r, is usually taken from experimental structural data obtained by X-ray crystallography [28–30] or nuclear magnetic resonance spectroscopy [27] whereas the initial velocities are assigned randomly according to the Maxwell-Boltzmann distribution such that the net momentum $\sum m_i \dot{r}_i$ is zero [116,123]. The Verlet and leapfrog integrators are commonly used. These second-order integrators are time-reversible and symplectic. Note that the derivatives of U are readily available because of its simple form.

Integration of Newton's equations of motion does not change the total energy E, i.e., the microcanonical ensemble (NVE, constant number of particles N, volume V, and energy E) is obtained. For simulations in the canonical ensemble (NVT, constant N, V, and temperature T) or isothermal-isobaric ensemble (NPT, constant N, pressure p, and T), e.g., for comparison with experimental data collected at the corresponding conditions, dedicated algorithms (thermostats and barostats) are available [132].

1.2.2 The time scale problem

The time step for the numerical integration of the equation of motion has to be sufficiently small to capture the fastest motions in the system and is typically set to about 1 fs [116,132]. The time step is thus several orders of magnitude smaller than the time scale of many biologically relevant processes, a fact that prohibits long simulations.

Several approaches exist to alleviate the time scale problem. First, it is very common to use algorithms like SHAKE [133] or LINCS [134] to constrain the lengths of covalent bonds involving hydrogen atoms. This essentially removes the fastest motions in the system, and a larger integration time step can be chosen. Second, coarse-graining provides an alternative means for reducing the number of degrees of freedom in the system [135–137]. A prominent example of coarse-graining is to model the solvent implicitly based on a continuum approximation [138–140]. In this case, computational cost is reduced significantly with respect to explicit solvent simulations that require a large

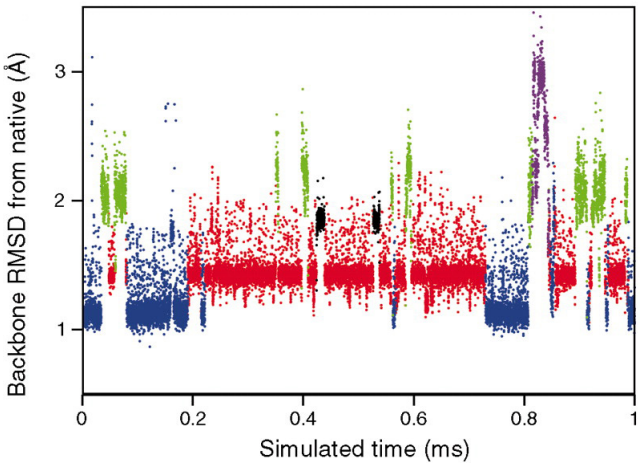


Figure 1.6: **Dynamics of folded BPTI in a 1 ms long MD simulation.** The root-mean-square deviation (RMSD) from a crystal structure (PDB ID 5PTI [215]) shows interconversion among different metastable states. The colours indicate the five metastable states identified by Shaw et al. [106]. Reprinted with permission from Ref. [106].

number of solvent-solvent interactions to be evaluated. Third, a variety of algorithms are available to enhance sampling of rare conformations and events [141–145]. Fourth, regarding hardware, MD simulations on distributed computing systems [6,146], special-purpose supercomputers [147,148] as well as on graphical processing units [149–151] have been reported.

Some of these approaches can be combined with each other, and the art of molecular simulation consists in choosing the appropriate model to answer a given question. Fast models with reduced resolution and accuracy might be sufficient if the purpose of a simulation is generation of hypotheses.

The first all-atom MD simulation of a small protein in explicit solvent reaching 1 ms was reported in 2010 [106]. Figure 1.6 shows that the 58-residue protein BPTI (Fig. 1.2c) interconverts among a handful of metastable states in this simulation. We use this data set in the following and in Chapters 3 and 5.

1.3 Analysis of molecular dynamics simulation data

In general, analysis of MD simulation data is concerned with charactering emerging behaviours observed in the simulation, i.e., with deriving macroscopic dynamics and with elucidating slow processes in terms of associated structural changes and rates [17,18,152–154]. An essential task is thus to understand the topography of the free energy surface. Free energy minima, i.e., metastable states, have to be detected and characterised, and the connectivities among them have to be described. For such analyses, which are often exploratory in practice, a wide range of algorithms exist, and these are reviewed and classified next.

1.3.1 Methods

Dimensionality reduction A wide range of algorithms has been developed to represent the original high-dimensional data in a space of lower dimensionality such that the most important aspects of the data set are preserved. Different algorithms are based on different mathematical translations of the ambiguous term "most important aspects", and this means that the appropriateness of dimensionality reduction algorithms is specific to the question.

The simple and widely-used principal component analysis (PCA) rotates and projects the data such that most of the variance in the data is retained while the individual dimensions are made to be uncorrelated. Specifically, the i-th principal component u_i is determined as the direction of maximal variance in the data under the condition that u_i is linearly uncorrelated to $u_1, \ldots, u_{i-1}$ [155]. This task can be formulated as an eigenvalue problem. Applied to MD data, PCA [156] and related methods like multidimensional scaling [157, 158] and local feature analysis [159–161] capture large-amplitude motions.

However, large-amplitude motions are often not the most relevant ones. Thus, other algorithms have been developed, e.g., full correlation analysis [162], a method that searches for maximally anharmonic dimensions, and time-structure based independent component analysis (tICA). The latter is a method formally similar to PCA that uses kinetic information [163–168]. More precisely, the i-th time-structure based independent component v_i is determined as the direction of maximal autocorrelation in the data (for a fixed lag time) under the condition that v_i is uncorrelated to $v_1, \ldots, v_{i-1}$. As in PCA, the directions are given by the solution of an eigenvalue problem. The method is very promising since the functionally most relevant motions are typically also among the slowest ones, at least on the time scales currently accessible by MD simulations.

Another method for dimensionality reduction are diffusion maps [169]. Here, the data set is treated as a weighted graph with edge weights given by a kernel function, i.e., a positive semi-definite, symmetric function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ that measures similarity among data points. A Markov process on this graph is defined based on the weight matrix. This emulates a random walk on the data, and a low-dimensional projection of the data is given through the slowest eigenfunctions of the Markov process. The main advantages with respect to the other dimensionality reduction algorithms mentioned above are that the method is nonlinear and, with an appropriate kernel function, local. The method is related to spectral clustering [170], and it has been applied to MD simulation data [161, 171–175].

After reducing dimensionality with one of these methods, the data is typically represented with a two-dimensional histogram, free-energy surface, or scatter plot (Fig. 1.7). More often than not, two dimensions are not sufficient to capture all the relevant aspects of the data set without introducing significant overlap of distinct states, and visual processing of the result is impossible if more than a few dimensions are important. Alternatively, the contribution of individual amino acids to the various new dimensions can be visualised on the protein structure, thus highlighting the structural transitions associated with the new dimensions [176]. Diffusion maps reduce dimensionality and at the same time provide a kinetic model for the data [177, 178].

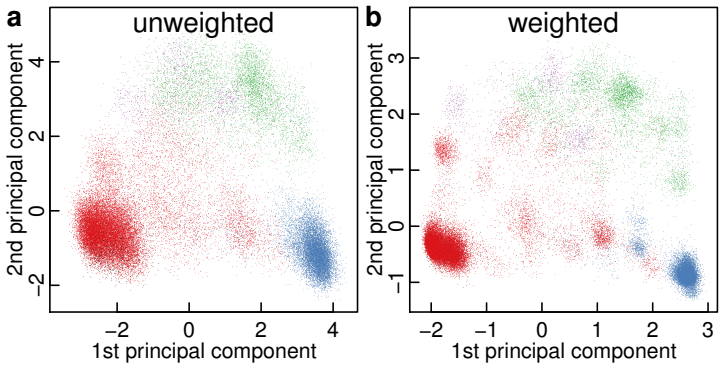


Figure 1.7: **Principal component analysis (PCA) for folded BPTI.** The data set originates from a MD simulation of the conformational dynamics of BPTI within its native state (Fig. 1.6) [106]. **(a)** Projection of the data (sine and cosine values of 271 nonsymmetric dihedral angles for 41250 snapshots) onto the first two principal components without prior scaling of the input features. **(b)** Projection of the data onto the first two principal components after scaling the input features according to their autocorrelation functions, thus enhancing the influence of slow dihedral angles (see Chapter 5 for details). The projection shown in (b) reveals more structure in the data compared to the one in (a). The data points are coloured as in Fig. 1.6.

We note that comparison of different dimensionality reduction techniques is notoriously difficult since no objective cost function exists [179], and since many new methods are only compared to PCA.

Clustering A second class of methods is based on clustering the data set, i.e., on grouping together similar conformations [8, 180]. Clustering is most often based purely on geometric criteria, but a method that incorporates kinetic information at the very first step has been devised [106]. The conformations of the system can be visualised directly for a coarse clustering consisting only of a handful of clusters (Fig. 1.8), but this is rarely possible without combining what are essentially distinct states.

The transitions between clusters that were observed in the original time series can be used to construct a network representation of the data [181]. These networks can be visualised with algorithms for creating graph layouts (Fig. 1.9) [67, 182], cut-based free energy profiles (Fig. 1.10) [183, 184], or free energy disconnectivity graphs [185–187]. Quantitative analysis with tools from graph theory can be performed [67, 188, 189], and investigations of kinetics in terms of discrete Markov models and transition-path theory [69, 190–192] are common. However, Markov models are difficult to parametrise and depend heavily on the clustering algorithm and the number of clusters [193]. Recently, hidden Markov models have been applied to MD data [194, 195], and several methods for coarse-graining high-resolution Markov models have been reported [196].

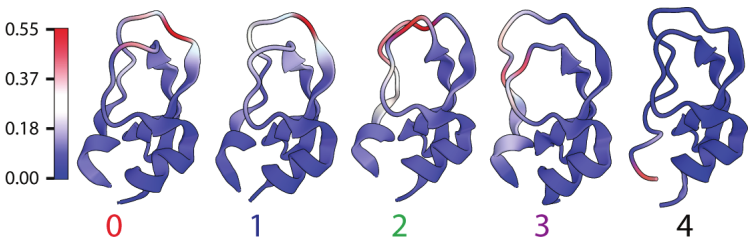


Figure 1.8: **Cluster analysis for folded BPTI.** Representative structures are shown for the five clusters identified by Shaw et al. in the trajectory presented in Figs. 1.6 and 1.7 [106]. Residues coloured in red tend to adopt a conformation that is unique for the given cluster (see Ref. [106] for details). Reprinted with permission from Ref. [106].

Variational approach A variational principle has been derived to approximate the slow processes captured by a master equation (Eq. 1.6) in discrete time [166,168,197,198]. This approach circumvents the data-driven discretisation of the state space discussed above. However, a set of basis functions has to be provided upfront, either based on intuition and prior knowledge of the system [166,198] or derived algorithmically [168].

Low-dimensional diffusion As mentioned above, protein folding in particular has been described as diffusion in a one-dimensional potential. The main challenge is to find an appropriate reaction coordinate [199]. While the number of native contacts can be a reasonably good reaction coordinate in some cases (as suggested by the metaphor of a funnelled free energy surface) [92], dedicated optimisation schemes were shown to give superior result in other applications [200–202]. Since it is currently not clear which processes can ultimately be modelled as simple one-dimensional diffusion problems, analysis tools imposing this theoretical framework are of limited general use. However, recent theoretical advances might extend their applicability [203].

The modelling of protein dynamics as diffusion in higher-dimensional spaces has been attempted [204–207]. However, this task is difficult since the position-dependent diffusion tensor has to be estimated. To be applicable and interpretable in practice, these methods rely heavily on prior dimensionality reduction. Therefore, they inherit a number of drawbacks, and these are discussed next.

1.3.2 Typical problems

The methods mentioned above all have several of the following drawbacks. First, many are not scalable to large data sets. For example, diffusion maps require all pairwise distances as input [169] and thus scale quadratically with data set size. Second, many of these methods do not preserve information, i.e., states do overlap or kinetic shortcuts are introduced (Fig. 1.11) [181,189,208]. Third, the heavy parameter dependence of some methods makes their use cumbersome, renders their outcome less objective, and can even impede parametrisation of the associated model for high-dimensional data. Fourth, the amount of information they provide can be rather limited.

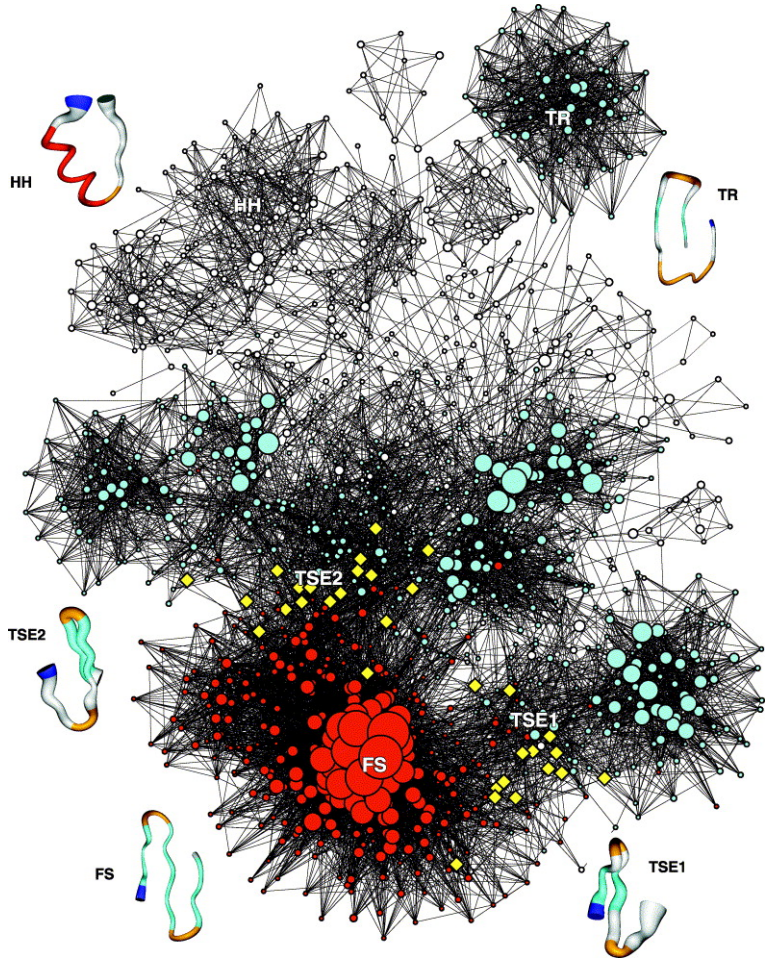


Figure 1.9: **Network analysis for the peptide Beta3S.** The underlying data is comprised of 10 μs of MD simulations during which the peptide folded and unfolded about 70 times [67]. Nodes in the network represent groups of snapshots with similar secondary structure. The transitions among snapshots observed in the simulation provide the links among the nodes. The size of the nodes reflects the number of corresponding snapshots. Nodes in red, cyan, and white have high, intermediate, and low connectivity to other nodes. The folded state (FS) of Beta3S can be reached through two structurally distinct transition state ensembles (TSE1 and TSE2, yellow nodes). The denatured state ensemble is heterogeneous and features helical conformations (HH) and low-enthalpy conformations that can act as traps (TR). Representative snapshots of selected nodes are shown as pipes, which are coloured according to secondary structure with the N-terminus in blue. The width of the pipe reflects structural heterogeneity within a node. Reprinted with permission from Ref. [67].

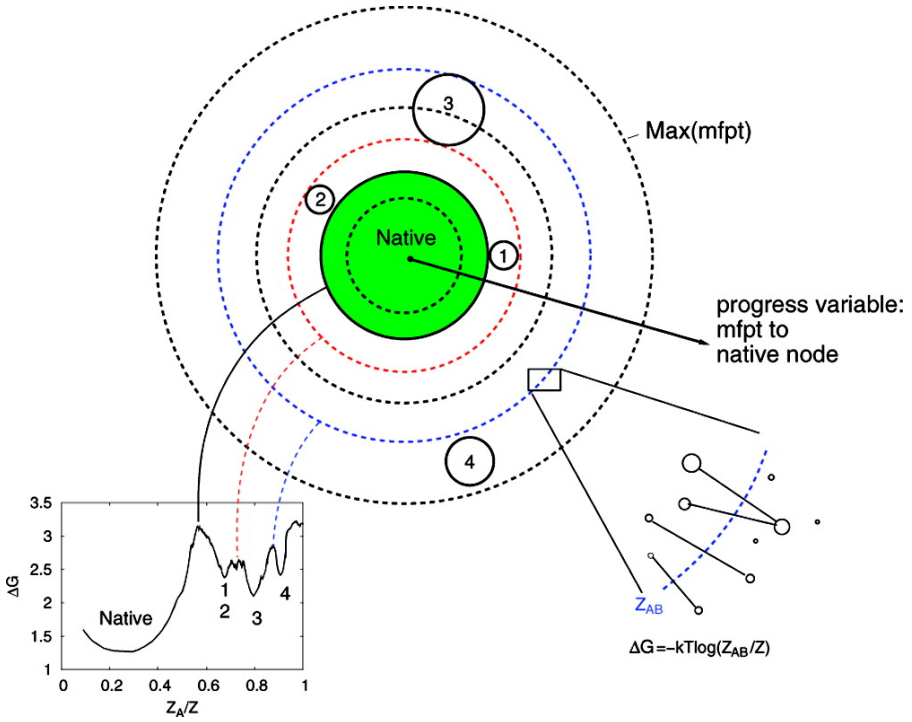


Figure 1.10: **Illustration of the cut-based free energy profile (cfep) procedure.** Cfeps provide a means to visualise network representations of MD data (Fig. 1.9) and are constructed as follows. First, a reference node is selected, typically a native conformation in the context of folding. Second, the network is treated as a discrete Markov model. This allows an analytical computation of the mean first-passage times to the reference node for the remaining nodes v (labelled 1–4 in the figure), which are denoted $mfpt_v$. Third, a point is drawn at $(x = Z_A/Z, y = \Delta G = -kT \log Z_{AB}/Z)$ for every node v to obtain the cfep (lower left panel). Here Z is the total number of snapshots in the data, Z_A is the cumulative size of all the nodes u with $mfpt_u < mfpt_v$, and Z_{AB}/Z is the number of transitions in the simulation between the nodes u with $mfpt_u < mfpt_v$ and the nodes w with $mfpt_w > mfpt_v$ (lower right diagram). The nodes are thus sorted along the x-axis according to their kinetic distance from the reference node. The free energy $\Delta G = -kT \log Z_{AB}/Z$ is high in bottleneck regions of the network (e.g., black, solid circle) and low elsewhere, which allows identification of free-energy basins. Distinct free energy basins with similar $mfpt$ to the reference state overlap in the cfep (nodes 1 and 2), which is an inherent drawback of the method. Reprinted with permission from Ref. [184]. Copyright (2008) American Chemical Society.

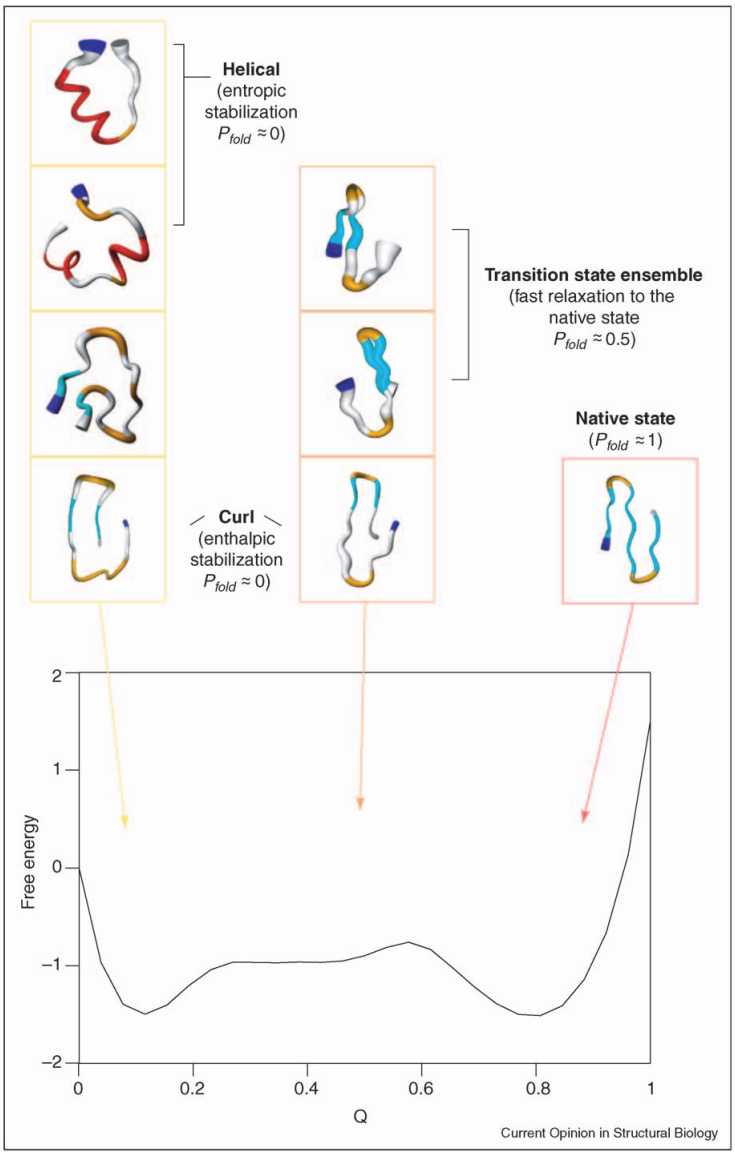


Figure 1.11: **Projections onto geometric order parameters can mask important information.** The free energy of Beta3S is plotted as a function of the fraction of native contacts (Q) for the data underlying Fig. 1.9. A few representative structures are shown (see Fig. 1.9 for details). Distinct states overlap on this projection, e.g., helical and curl-like conformations are both found at the left end of the profile (yellow boxes). Moreover, trapped conformations (bottom orange box), which are kinetically far from the folded state (red box), can have up to half of the native contacts formed. These conformations thus overlap with structures from the transition state ensemble (top two orange boxes), i.e., the order parameter Q does not reflect kinetic distance. Reprinted with permission from Ref. [181].

We have developed and tested an analysis algorithm not suffering from these drawbacks. The main output of this method is the Sapphire plot, which is presented in the next subsection.

Finally, almost all of the methods, including Sapphire plots, depend on how the input data is represented or how distance between snapshots is measured (Fig. 1.7). An informative representation of the data is often obtained with the help of prior knowledge, and results will be biased by this sometimes subjective choice. This is especially relevant for linear methods and for algorithms that rely solely on geometric information. Potential solutions for this issue as well as our ongoing work are discussed below.

1.4 Sapphire plots

The main contribution of the present work is a novel method to analyse and visualise MD simulation data. The proposed method generates a Sapphire (States And Pathways Projected with HIgh REsolution) plot, which represents the states and the sequence of events sampled by the simulation. Specifically, the snapshots are arranged linearly such that snapshots belonging to the same high-density region are grouped together. The resultant sequence of snapshots is called progress index and serves as the x-axis in the Sapphire plot. Then, the progress index is annotated in various ways to highlight, distinguish, and characterise the different high-density regions in the data.

We briefly introduce and discuss Sapphire plots here and refer the reader to the original publication [22], which is provided in Chapter 2, for detailed information. Applications to protein folding, to protein dynamics and to peptide binding are treated in Chapters 3–5.

1.4.1 The progress index

A function measuring distance between snapshots is required to construct the progress index, i.e., to sort the snapshots. The distance function can be chosen freely. We emphasise that the results depend on this choice. Indeed, finding appropriate distance functions for unsupervised learning from MD simulation data is an active area of research (see below and Chapter 5).

Once a distance function has been defined, all the snapshots are arranged sequentially in stepwise fashion starting from an arbitrary snapshot. In each step, the snapshot closest to any snapshot prior in the progress index becomes the next entry.

The crucial property of this walk through the data is that snapshots from the same high-density region are grouped together, and that distinct states do not overlap [22]. Free-energy basins can thus be identified along the progress index if the distance function is chosen such that these basins correspond to regions of high sampling density. We point out that the progress index is formally similar to the output of the Optics algorithm [209].

It is important to note that the progress index is not meant to serve as a reaction coordinate. Furthermore, the data set is not required to come from a single trajectory.

Multiple trajectories of different lengths can be combined.

The following example uses a two-dimensional data set to illustrate the progress index.

Illustration with the Müller potential The Müller potential [210] is an illustrative, two-dimensional model potential that has been used to test algorithms to compute reaction paths [192, 210–212] as well as analysis tools for MD simulation data [168, 193]. It is given by $V(r = (x, y)) = \sum_{i=1}^{4} p_i \exp(a(x - \bar{x}_i)^2 + b(x - \bar{x}_i)(y - \bar{y}_i) + c(y - \bar{y}_i)^2)$ with $p = (-200, -100 - 170, 15)$, $a = (-1, -1, -6.5, 0.7)$, $b = (0, 0, 11, 0.6)$, $c = (-10, -10, -6.5, 0.7)$, $\bar{x} = (1, 0, -0.5, -1)$, and $\bar{y} = (0, 0.5, 1.5, 1)$. Fig. 1.12a shows the contour lines of the potential highlighting three distinct minima.

Diffusion on the Müller potential was modelled by the Langevin equation (Eq. 1.10) $\mathrm{d}r/\mathrm{d}t = -\nabla V(r)\zeta + \sqrt{2k_BT\zeta}R(t)$ where $R(t)$ is a delta-correlated Gaussian process with zero mean, and the unitless parameters were set to $k_BT = 20$ and $\zeta = 10^{-3}$.

The numerical simulation was performed with the Euler-Maruyama method [213], i.e., $r_0 = (1, 0)$ and $r_{t+1} = r_t - \nabla V(r_t)\zeta\Delta t + \sqrt{2k_BT\zeta\Delta t}z_t$ where $z_t \sim \mathcal{N}(0, 1)$. The time step Δt was set to 0.1 with a total simulation length of 10^5 steps. Coordinates were saved every 10 steps, and the simulated trajectory is visualised in Fig. 1.12b. We observed about a dozen transitions into the major basin at $(-0.5, 1.5)$.

We computed the progress index for this data set starting from the first snapshot in the trajectory and using the Euclidean distance function. Figure 1.12b visualises the progress index and shows that it indeed traces all three minima one by one.

Two characteristics should be noted here. First, snapshots found in the inner and shared boundary region of the two minima with high x and low y values (coloured orange to yellow in Fig. 1.12b) are grouped together by construction. The same observation applies to the snapshots in the outer boundary regions surrounding all three minima (coloured purple in Fig. 1.12b). This happens independently of sampling density and can potentially be avoided by adapting the distance function. Second, once the progress index enters a newly encountered high-density region (e.g., around progress index values of about 2500 in Fig. 1.12b) it proceeds through a "narrow tube" (yellow snapshots around (-0.8,1.2)) towards the centre of maximum density (green snapshots). After this point is reached, remaining snapshots from the same high-density region are added, thus filling the basin concentrically (snapshots coloured turquoise to blue).

We have observed in some applications that the progress index can prematurely enter new basins [23]. This can be rectified by increasing the sampling density.

Scalable approximate algorithm We have developed a stochastic algorithm to generate an approximate progress index [22]. This algorithm is scalable to large data sets and is based on the observation that the progress index is related to a minimum spanning tree (MST) constructed on the data points. To see this, let $G = (V, E)$ be the complete graph with nodes V given by the snapshots and edges $e = \{u, v\} \in E$ weighted by $w_e = d(u, v)$, the distance between snapshots u and v. A MST of G is a subgraph $T = (V, E')$ of G that connects all nodes and whose total weight $\sum_{e \in E'} w_e$ is minimal. If a MST is known, the progress index can be derived in $\mathcal{O}(N \log N)$ time, where N is

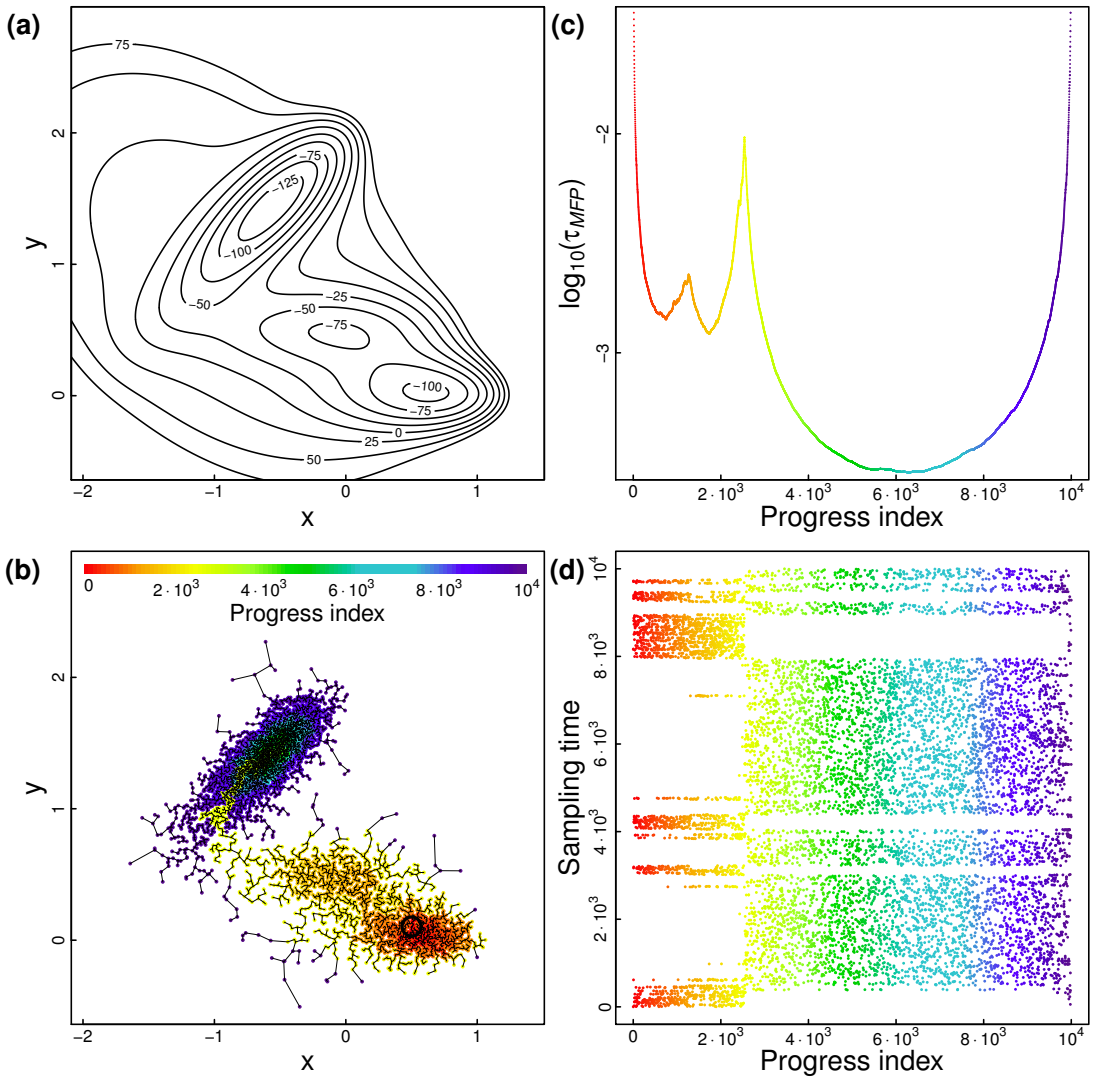


Figure 1.12: **Sapphire plot for the Müller potential.** **(a)** Contour lines of the Müller potential. **(b)** Simulated trajectory. Data points are coloured according to the progress index, as indicated on top, and the first snapshot in the progress index is marked with a black circle. A line is drawn between every snapshot u and its parent in the progress index, i.e., the snapshot closest to u among those added to the progress index before u. These connections form a minimum spanning tree on the data (see text). **(c)** Kinetic annotation of the progress index revealing the three minima of the potential. Data points are coloured according to the progress index, as indicated in (b). **(d)** Dynamic trace on the progress index, i.e., annotation with sampling time. Data points are coloured according to the progress index, as indicated in (b).

the number of snapshots. Conversely, if we define E' as the set of edges between every snapshot u and its parent in the progress index, i.e., the snapshot closest to u among those added to the progress index before u, then $T = (V, E')$ is a MST of G (Fig. 1.12b).

The approximate stochastic algorithm that is scalable to large data sets first organises the data with a hierarchical clustering [180] and then uses heuristics to compute a short spanning tree (SST) S of G, i.e., a spanning tree whose total weight is small but not necessarily minimal. While the clustering is deterministic, the SST is stochastic. The SST S is then used to derive an approximate progress index, as described in Chapter 2.

1.4.2 The annotation functions

With the progress index in hand, annotation functions are needed to highlight and interpret the states along the progress index and the pathways connecting them. To this end, we use the following three different types of annotation functions.

Kinetic annotation function The first annotation function indicates boundaries between states along the progress index in a kinetic sense and is constructed as follows. For every snapshot i along the progress index, we compute the average of the mean first-passage times between A_i and S_i, where A_i is the set of snapshots added to the progress index before i and S_i is the set of those added after i. This quantity, denoted τ_{MFP}, can be computed analytically and efficiently. The value of τ_{MFP} is low within a state and high in transition regions [22].

Figure 1.12c shows τ_{MFP} for the illustration based on the Müller potential, and the three basins are clearly highlighted.

Dynamic trace The second annotation is the sampling time of the individual snapshots and illustrates when the different states were encountered. This information traces the temporal evolution of the system and is thus called the dynamic trace.

For the illustration based on the Müller potential the dynamic trace is given in Fig. 1.12d.

The dynamic trace is particularly useful to quickly asses recurrence, i.e., whether a given state was sampled multiple times. SAPPHIRE plots can thus serve as a tool to check for convergence in the simulation [144]. Furthermore, the dynamic trace visualises the detailed sequence of events and can inspire mechanistic models of a process like peptide binding (Chapter 4).

Structural annotation In addition, we characterize the states structurally. Possible annotation functions include a secondary structure assignment by residue [214], similarity to a reference snapshot, intermolecular and intramolecular distances, dihedral angles, and data from other analysis methods. The structural annotations naturally show the structural heterogeneity within states since every single snapshot is represented by the progress index, i.e., since resolution is maximal.

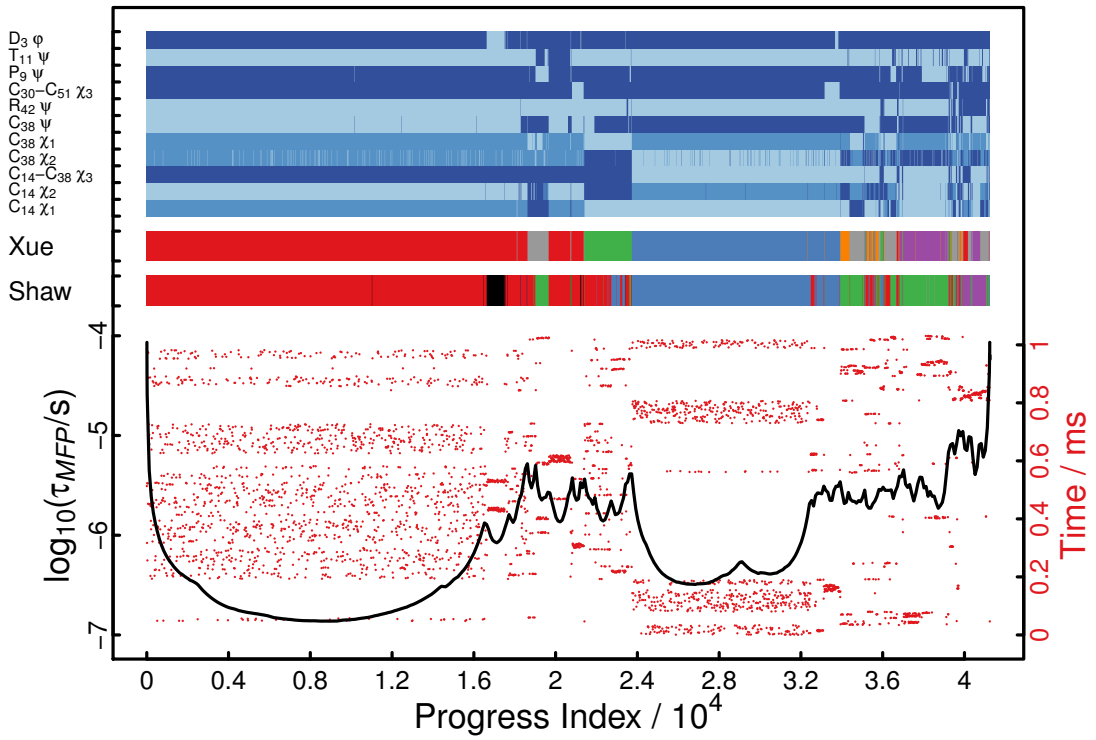


Figure 1.13: **Sapphire plot for BPTI.** The underlying trajectory is the same as in Figs. 1.6 and 1.7. The progress index (x-axis) is annotated with kinetic information (black curve), dynamic trace (red dots), and structural information (color annotation on top). The latter includes colour-coded state assignments according to Shaw et al. [106] (Fig. 1.8) and Xue et al. [217] (M1 - blue, M2 - orange, M3 - magenta, mC14 - red, mC38 - green, and other states - grey). The colour-coded information on top uses binning of selected dihedral angles and shows that states differ in their arrangement of the Cys14-Cys38 and Cys30-Cys51 disulfide bonds and of the N-terminal part of the protein. See Fig. 6b in Chapter 5 for details.

1.4.3 Application to BPTI

We now touch upon an application of Sapphire plots to MD simulation data [106] of the native state ensemble of the 58-residue protein BPTI (Fig. 1.6). Here, we merely point out the wealth of information contained in a Sapphire plot and refrain from discussing the biophysical relevance of this data set and our findings since the very same data is treated comprehensively in Chapters 3 and 5, and since the associated literature is surveyed there.

The trajectory, which consists of 41250 snapshots saved every 25 ns, reveals that BPTI interconverts between several states on the μs time scale (Fig. 1.6) [106]. Figure 1.13 shows a Sapphire plot for this trajectory based on a weighted Euclidean distance function of the dihedral angles of BPTI (see Chapter 5) and confirms the presence of

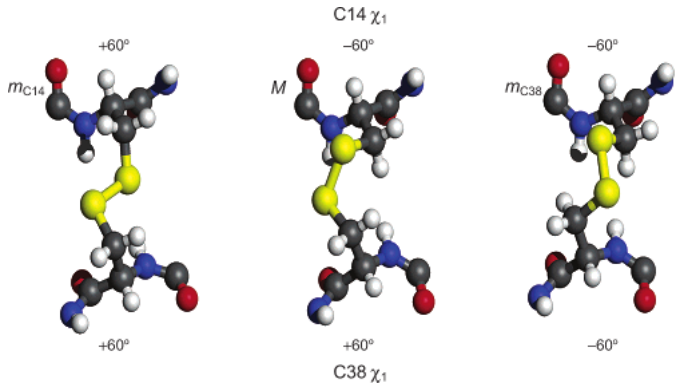


Figure 1.14: **Metastable conformations of the Cys14-Cys38 disulfide bond of BPTI.** The model is based on NMR data and encompasses a major (M) and two minor (m_{C14} and m_{C38}) states. The average values for the χ_1 dihedral angles of Cys14 and Cys38 are indicated above and below each conformation. Xue et al. have extended this model by splitting M into three states [217], and we have used their model in Fig. 1.13. Reprinted with permission from Ref. [216]. Copyright (2003) American Chemical Society.

several metastable states.

The structural annotation demonstrates that most of these states can be characterised by different conformations of the disulfide bonds, and there is experimental evidence for several of them (Fig. 1.14) [32,215,216]. The dynamic trace unmasks when and in which sequence these states were encountered. In particular, it shows that the state found at progress index values of about 24000 to 32500, which corresponds to the crystal structure (PDB ID 5PTI) [215], is sampled multiple times while the state at the very end of the progress index (purple in the annotation by Shaw et al.) is sampled only once. The latter observation indicates a lack of recurrence.

Overall, Fig. 1.13 shows that the picture emerging from the Sapphire plot agrees very well with the clustering into five states reported in the original publication of the data by Shaw et al. [106] but is richer in details, and the same holds in comparison with the model of Xue et al. [217], which is mainly based on domain knowledge.

1.4.4 Conclusion

The Sapphire plot is an excellent visual tool for exploratory data analysis and communication of results in a single plot. It delineates the individual metastable states of a complex system like a biomolecule, and thermodynamic information, namely the sampling weights of individual states, can be resolved quantitatively. In terms of kinetics, pathways and the sequence of events are visualised in high detail, and states can readily be defined to compute rates or transition path times [23]. Furthermore, the Sapphire plot enables rapid assessment of state interconversion and recurrence, i.e., it provides information on sampling convergence.

The Sapphire plot does not rely on any parameters controlling size, number, shape

or other properties of states. Distinct states do not overlap on the progress index, in contrast to projections using geometric or kinetic distances from a reference state to order snapshots (Figs. 1.10 and 1.11) [181,189,208]. Moreover, every snapshot is represented on the progress index, i.e., resolution is maximal, and the structural annotation functions characterise state-specific conformations as well as any variety therein. This is in contrast to methods that rely on binning, e.g., two-dimensional projections or clustering (Fig. 1.8).

The progress index can be computed in $\mathcal{O}(N \log N)$ time with an approximate, stochastic algorithm whereas the annotation functions scale as $\mathcal{O}(N)$. The method is thus scalable to large data sets, a substantial advantage over algorithms relying on the full distance matrix like multidimensional scaling [157,158] or diffusion maps [169].

The algorithm to construct the progress index and to compute the kinetic annotation function is implemented in the CAMPARI simulation and analysis package, which is freely available (http://campari.sourceforge.net).

SAPPHIRE plots have been used to investigate data from MD simulations of protein folding [22,23] (Chapters 2, 3, and 5), of the conformational dynamics of proteins [23] (Chapters 3 and 5), of a loop of the prion protein [218], and of peptide binding [219] (Chapter 4). Finally, we mention that the progress index is being used as a means to guide sampling [145].

1.5 Distance functions

As mentioned above and hinted at in Fig. 1.7, the outcome of most analysis algorithms for MD simulation data depends on how the data is represented or how distance between snapshots is quantified. If the slow processes of a complex system are the main topic of interest, the kinetic proximity of two snapshots would be an ideal distance function. This quantity, however, is not accessible in practice. Structural distance functions act as proxies in data analysis with the hope that structural similarity according to the chosen metric reflects kinetic distance. We introduce the most common choices for such distance functions next and refer the reader to the literature for others [220,221].

RMSD The coordinate root-mean-square deviation (RMSD) between two conformations x and y, which is the most widely used distance function, is defined as

$$d_{RMSD}(x,y)^2 = \min_{R,t} \frac{1}{D} \sum_{i=1}^{D} \|r_{x,i} - R r_{y,i} - t\|^2, \tag{1.13}$$

where D is the number of atoms considered, $r_{x,i}$ is the three-dimensional vector containing the coordinates of atom i in conformation x, R is a 3×3 rotation matrix, and $t \in \mathbb{R}^3$. Thus, R and t define an optimal alignment of x and y.

dRMSD The distance root-mean-square deviation (dRMSD) is based on a list of D interatomic distances and defined as

$$d_{dRMSD}(x,y)^2 = \frac{1}{D}\sum_{i=1}^{D}(d_{x,i} - d_{y,i})^2, \tag{1.14}$$

where $d_{x,i}$ is the i-th interatomic distance in conformation x. Note that selecting all $(N-1)N/2$ possible atom pairs, where N is the number of atoms, results in an artificially increased dimensionality since the system has only $3N-6$ internal degrees of freedom.

A discretised version of d_{dRMSD} is obtained by replacing the interatomic distances by $H(c-d_{x,i})$, where H is the Heaviside function and c is a cutoff.

Dihedral angles Alternatively, distance can be measured based on a list of D dihedral angles, i.e.,

$$\begin{aligned} d_{dihed}(x,y)^2 &= \frac{1}{2D}\sum_{i=1}^{D}(\sin\alpha_{x,i} - \sin\alpha_{y,i})^2 + (\cos\alpha_{x,i} - \cos\alpha_{y,i})^2 \\ &= \frac{1}{D}\sum_{i=1}^{D} 1 - \cos(\alpha_{x,i} - \alpha_{y,i}), \end{aligned} \tag{1.15}$$

where $\alpha_{x,i}$ is the i-th dihedral angle in conformation x. This distance function neglects degrees of freedom like bond lengths and angles, and it cannot capture the distance between multiple molecules and their relative orientation.

Since biomolecular systems exhibit symmetries, e.g., indistinguishable water molecules and rotational symmetries in side chains, only nonsymmetric atoms or dihedral angles are typically used in Eqs. 1.13, 1.14, and 1.15.

Distance functions for high-dimensional data The distances defined in Eqs. 1.13, 1.14, and 1.15 require a selection of which atoms, interatomic distances, or dihedral angles to use. The full set of features usually contains a large number of irrelevant features that are not related to the processes of interest and can mask the relevant ones (Chapter 5). Further complications arise if the importance of different features varies among the data points [222–224]. As a consequence, choosing a suitable distance function can be a time-consuming task more important than the actual analysis method itself [8, 16, 19, 225]. This calls for efficient protocols to derive distance functions that reduce the influence of irrelevant features and account for local feature relevance.

1.5.1 Data-driven distance functions

We give here are short overview of the literature on data-driven methods to define distance functions. A summary of our proposed method for MD data concludes the section.

Feature selection and extraction For high-dimensional data, it is common to select or generate informative features, a process that often relies on domain knowledge or measures of relevance, such as entropy or mutual information [226]. A less drastic alternative to feature selection is feature weighting [16], i.e., using a weighted distance function of the form

$$d(x,y)^2 = \left(\sum_{i=1}^{D} w_i\right)^{-1} \sum_{i=1}^{D} w_i (a_{x,i} - a_{y,i})^2, \tag{1.16}$$

where $a_{x,i}$ is the value of the i-th feature for snapshot x, and w_i is the weight associated with the i-th feature. For d_{RMSD}, the resultant weighted alignment problem can be solved easily [227].

The methods for dimensionality reduction reviewed above are routinely used to extract features in the form of low-dimensional embeddings, although any low-dimensional embedding might be of limited use if the original data contain many irrelevant features and the distance function implied by the data space is unable to distinguish between similar and dissimilar points. Several variants of classical dimensionality reduction algorithms have been motivated by the latter observation. Isomap [179, 228–230] and sketch-map [231], for example, apply multidimensional scaling [157] to non-Euclidean input distances, and locally scaled diffusion map is an extension of diffusion maps [169] using a Gaussian kernel with snapshot-dependent local scales [173]. In any case, these methods introduce a choice of how many dimensions to keep.

In contrast to many other dimensionality reduction algorithms, tICA generates a low dimensional embedding that is invariant to linear, invertible transformations of the input data. In its basic form, the method is global and linear, but kernel-based extensions might allow to capture nonlinear structure in the data [168, 232].

Metric learning If pairs of snapshots are given that are defined to be similar, distance functions can be learned that reproduce these relationships [233]. In the context of MD simulation data, this approach has been used to learn a distance function that tends to return low values for pairs of data points that are close in time along the trajectory [234]. This task was formulated as a complex optimization problem depending on several parameters.

Locally adaptive semimetrics To account for local feature relevance, locally adaptive similarity measures have been used in classification [16, 235] and clustering [223, 224, 236–238].

The COSA (clustering objects on subsets of attributes) algorithm [236], for example, returns a locally adaptive semimetric, i.e., a symmetric function d satisfying $d(x,y) = 0 \iff x = y$, given by

$$d(x,y)^2 = \left(\sum_{i=1}^{D} \max(w_{x,i}, w_{y,i})\right)^{-1} \sum_{i=1}^{D} \max(w_{x,i}, w_{y,i})(a_{x,i} - a_{y,i})^2. \tag{1.17}$$

Here, every data point x has an associated set of weights $\{w_{x,i}\}_{i=1}^{D}$ that reflect the variance along the individual features among the nearest neighbours of x (determined according to Eq. 1.17).

For temporal data Singer et al. proposed the semimetric

$$d(x,y)^2 = (a_x - a_y)^\top \left(\Sigma_x^{-1} + \Sigma_y^{-1}\right)(a_x - a_y)\,, \tag{1.18}$$

where Σ_x is a local covariance matrix associated with x [239,240]. It can be determined by running short stochastic simulations starting from x [239,240] or from the data within a short time window along the trajectory around x [241]. This semimetric has been used in conjunction with diffusion maps for dimensionality reduction [240].

A locally adapted version of Eq. 1.15 was proposed specifically for MD simulation data of proteins [180]. The contribution of individual dihedral angles is weighted with the sum of the associated, snapshot-dependent moments of inertia. This semimetric was presented in the context of clustering and seamlessly integrated with the implementation of the clustering algorithm. Such a definition of weights is physically motivated but faces problems since information about side chains becomes negligible for big enough peptides and proteins.

Our contribution In Chapter 5, we propose to globally or locally weight features based on effective rates, thus enhancing the influence of slow degrees of freedom. Global weights are defined based on the autocorrelation function, while locally adaptive weights reflect transition rates within a time window along the trajectory. Our approach can be readily combined with classical algorithms for dimensionality reduction or clustering. We extensively test our data-driven approach in conjunction with several unsupervised learning protocols on an illustrative model system and two MD data sets. Both feature weighting methods reveal slow side-chain dynamics within the native state of the peptide Beta3S. For BPTI, Sapphire plots employing weighted distance functions reveal the metastable state m_{C38} (green in the annotation according to Xue et al. in Fig. 1.13). This state, for which there is experimental evidence (Fig. 1.14), was not resolved in RMSD-based Sapphire plots or in the clustering of Shaw et al. (Fig. 1.8).

Bibliography

[1] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

[2] N. M. Ball and R. J. Brunner. Data mining and machine learning in astronomy. *Int. J. Mod. Phys. D*, 19(7):1049–1106, 2010.

[3] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[4] B. Berger, J. Peng, and M. Singh. Computational solutions for omics data. *Nat. Rev. Genet.*, 14(5):333–346, 2013.

[5] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U.S.A.*, 110(15):5915–5920, 2013.

[6] K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.*, 6(1):15–21, 2014.

[7] L. Einav and J. Levin. Economics in the age of big data. *Science*, 346(6210):1243089, 2014.

[8] A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.*, 31(8):651–666, 2010.

[9] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philos. Trans. R. Soc., A*, 367(1906):4361–4383, 2009.

[10] J. R. Kettenring. Coping with high dimensionality in massive datasets. *Wiley Interdiscip. Rev. Comput. Stat.*, 3(2):95–103, 2011.

[11] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE T. Vis. Comput. Gr.*, 19(3):495–513, 2013.

[12] T. Walter, D. W. Shattuck, R. Baldock, M. E. Bastin, A. E. Carpenter, S. Duce, J. Ellenberg, A. Fraser, N. Hamilton, S. Pieper, M. A. Ragan, J. E. Schneider, P. Tomancak, and J.-K. Hériché. Visualization of image data from cells to organisms. *Nat. Methods*, 7(3s):S26–S41, 2010.

[13] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A. C. Gavin. Visualization of omics data for systems biology. *Nature Met.*, 7(3s):S56–S68, 2010.

[14] C. Chen. Information visualization. *Wiley Interdiscip. Rev. Comput. Stat.*, 2(4):387–403, 2010.

[15] S. J. Rysavy, D. Bromley, and V. Daggett. DIVE: A graph-based visual-analytics framework for big data. *IEEE Comput. Graph. Appl.*, 34(2):26–37, 2014.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer, New York, 2nd edition, 2009.

[17] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten. Challenges in protein-folding simulations. *Nat. Phys.*, 6(10):751–758, 2010.

[18] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande. To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.*, 23(1):58–65, 2013.

[19] D. de Ridder, J. de Ridder, and M. J. T. Reinders. Pattern recognition in bioinformatics. *Briefings Bioinf.*, 14(5):633–647, 2013.

[20] M. Vendruscolo and C. M. Dobson. Protein dynamics: Moore's law in molecular biology. *Curr. Biol.*, 21(2):R68–R70, 2011.

[21] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw. Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophys.*, 41:429–452, 2012.

[22] N. Blöchliger, A. Vitalis, and A. Caflisch. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.*, 184(11):2446–2453, 2013.

[23] N. Blöchliger, A. Vitalis, and A. Caflisch. High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.*, 4:6264, 2014.

[24] D. Voet and J. G. Voet. *Biochemistry.* John Wiley & Sons, Inc., Hoboken, 4th edition, 2011.

[25] L. Tskhovrebova and J. Trinick. Titin: Properties and family relationships. *Nat. Rev. Mol. Cell Biol.*, 4(9):679–689, 2003.

[26] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[27] Kurt Wüthrich. *NMR of proteins and nucleic acids*. Wiley, New York, 1986.

[28] M. F. C. Ladd and R. A. Palmer. *Structure determination by X-ray crystallography*. Springer, New York, 5th edition, 2013.

[29] E. F. Garman. Developments in X-ray crystallographic structure determination of biological macromolecules. *Science*, 343(6175):1102–1108, 2014.

[30] Y. Shi. A glimpse of structural biology through X-ray crystallography. *Cell*, 159(5):995–1014, 2014.

[31] D. Fass. Disulfide bonding in protein biophysics. *Annu. Rev. Biophys.*, 41:63–79, 2012.

[32] G. Otting, E. Liepinsh, and K. Wüthrich. Disulfide bond isomerization in BPTI and BPTI(G36S): An NMR study of correlated mobility in proteins. *Biochemistry*, 32(14):3571–3582, 1993.

[33] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[34] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

[35] P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

[36] R. Das and D. Baker. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[37] C. Levinthal. How to fold graciously. *Proc. Mössbauer Spectrosc. Biol. Syst.*, pages 22–24, 1969.

[38] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.*, 89(1):20–22, 1992.

[39] K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. The protein folding problem: When will it be solved? *Curr. Opin. Struct. Biol.*, 17(3):342–346, 2007.

[40] R. F. Service. Problem solved* (*sort of). *Science*, 321(5890):784–786, 2008.

[41] P. G. Wolynes, W. A. Eaton, and A. R. Fersht. Chemical physics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 109(44):17770–17771, 2012.

[42] K. A. Dill and J. L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[43] D. M. Mitrea and R. W. Kriwacki. Regulated unfolding of proteins in signaling. *FEBS Lett.*, 587(8):1081–1088, 2013.

[44] T. Inobe and A. Matouschek. Paradigms of protein degradation by the proteasome. *Curr. Opin. Struct. Biol.*, 24:156–164, 2014.

[45] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[46] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, 2007.

[47] G. Bhabha, J. Lee, D. C. Ekiert, J. Gam, I. A. Wilson, H. J. Dyson, S. J. Benkovic, and P. E. Wright. A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, 332(6026):234–238, 2011.

[48] D. R. Glowacki, J. N. Harvey, and A. J. Mulholland. Taking Ockham's razor to enzyme dynamics and catalysis. *Nat. Chem.*, 4(3):169–176, 2012.

[49] K. Świderek, J. J. Ruiz-Pernía, V. Moliner, and I. Tuñón. Heavy enzymes - experimental and computational insights in enzyme dynamics. *Curr. Opin. Chem. Biol.*, 21:11–18, 2014.

[50] R. G. Smock and L. M. Gierasch. Sending signals dynamically. *Science*, 324(5924):198–203, 2009.

[51] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.

[52] O. G. Berg and P. H. von Hippel. Diffusion-controlled macromolecular interactions. *Annu. Rev. Biophys. Biophys. Chem.*, 14:131–158, 1985.

[53] G. Schreiber, G. Haran, and H.-X. Zhou. Fundamental aspects of protein-protein association kinetics. *Chem. Rev.*, 109(3):839–860, 2009.

[54] C. Tang, J. Iwahara, and G. M. Clore. Visualization of transient encounter complexes in protein-protein association. *Nature*, 444(7117):383–386, 2006.

[55] R. J. Ellis. Macromolecular crowding: An important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.*, 11(1):114–119, 2001.

[56] Y. Phillip and G. Schreiber. Formation of protein complexes in crowded environments - from in vitro to in vivo. *FEBS Lett.*, 587(8):1046–1052, 2013.

[57] G. E. Schulz and R. H. Schirmer. *Principles of protein structure*. Springer, New York, 1979.

[58] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand. Conformational entropy in molecular recognition by proteins. *Nature*, 448(7151):325–329, 2007.

[59] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, 5(11):789–796, 2009.

[60] R. Baron and J. A. McCammon. Molecular recognition and ligand association. *Annu. Rev. Phys. Chem.*, 64:151–175, 2013.

[61] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, 12(1):54–60, 2002.

[62] K.-B. Wong, T.-H. Yu, and C.-H. Chan. 3.2 Energetics of protein folding. *Compr. Biophys.*, 3:19–33, 2012.

[63] C. N. Pace. Energetics of protein hydrogen bonds. *Nat. Struct. Mol. Biol.*, 16(7):681–682, 2009.

[64] C. N. Pace. Conformational stability of globular proteins. *Trends Biochem. Sci.*, 15(1):14–17, 1990.

[65] R. Zwanzig. Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci. U.S.A.*, 94(1):148–150, 1997.

[66] C. M. Dobson, A. Šali, and M. Karplus. Protein folding: A perspective from theory and experiment. *Angew. Chem., Int. Ed.*, 37(7):868–893, 1998.

[67] F. Rao and A. Caflisch. The protein folding network. *J. Mol. Biol.*, 342(1):299–306, 2004.

[68] D. L. Ensign, P. M. Kasson, and V. S. Pande. Heterogeneity even at the speed limit of folding: Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.*, 374(3):806–816, 2007.

[69] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.*, 106(45):19011–19016, 2009.

[70] F. Liu, C. Dumont, Y. Zhu, W. F. DeGrado, F. Gai, and M. Gruebele. A one-dimensional free energy surface does not account for two-probe folding kinetics of protein α_3D. *J. Chem. Phys.*, 130(6):061101, 2009.

[71] S. V. Solomatin, M. Greenfeld, S. Chu, and D. Herschlag. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature*, 463(7281):681–684, 2010.

[72] M. Pirchi, G. Ziv, I. Riven, S. S. Cohen, N. Zohar, Y. Barak, and G Haran. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat. Commun.*, 2:493, 2011.

[73] J. Stigler, F. Ziegler, A. Gieseke, J. C. M. Gebhardt, and M. Rief. The complex folding network of single calmodulin molecules. *Science*, 334(6055):512–516, 2011.

[74] A. Fersht. *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding.* W. H. Freeman and Company, New York, 1999.

[75] R. L. Baldwin. How does protein folding get started? *Trends Biochem. Sci.*, 14(7):291–294, 1989.

[76] M. Karplus and D. L. Weaver. Protein-folding dynamics. *Nature*, 260:404–406, 1976.

[77] M. Karplus and D. L. Weaver. Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci.*, 3(4):650–668, 1994.

[78] D. B. Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 70(3):697–701, 1973.

[79] A. R. Fersht. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, 7(1):3–9, 1997.

[80] D. E. Makarov and K. W. Plaxco. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.*, 12(1):17–26, 2003.

[81] V. A. Voelz and K. A. Dill. Exploring zipping and assembly as a protein folding principle. *Proteins: Struct., Funct., Bioinf.*, 66(4):877–888, 2007.

[82] G. C. Rollins and K. A. Dill. General mechanism of two-state protein folding kinetics. *J. Am. Chem. Soc.*, 136(32):11420–11427, 2014.

[83] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.

[84] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. T. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young. Protein states and proteinquakes. *Proc. Natl. Acad. Sci. U.S.A.*, 82(15):5000–5004, 1985.

[85] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Struct., Funct., Bioinf.*, 21(3):167–195, 1995.

[86] R. Zwanzig. Diffusion in a rough potential. *Proc. Natl. Acad. Sci. U.S.A.*, 85(7):2029–2030, 1988.

[87] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267(5204):1619–1620, 1995.

[88] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, 14(1):70–75, 2004.

[89] M. Karplus. Behind the folding funnel diagram. *Nat. Chem. Biol.*, 7(7):401–404, 2011.

[90] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 84(21):7524–7528, 1987.

[91] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48(1):545–600, 1997.

[92] R. B. Best, G. Hummer, and W. A. Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.*, 110(44):17874–17879, 2013.

[93] B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia, and J. Clarke. Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature*, 463(7281):685–688, 2010.

[94] D. U. Ferreiro, E. A. Komives, and P. G. Wolynes. Frustration in biomolecules. *Q. Rev. Biophys.*, 47(4):285–363, 2014.

[95] N. G. Van Kampen. *Stochastic processes in physics and chemistry*. Elsevier, Amsterdam, 1992.

[96] P. Hänggi, P. Talkner, and M. Borkovec. Reaction-rate theory: Fifty years after Kramers. *Rev. Mod. Phys.*, 62(2):251–341, 1990.

[97] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.*, 104(15):5860–5868, 1996.

[98] B. Schuler, E. A. Lipman, and W. A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(6908):743–747, 2002.

[99] J. Kubelka, J. Hofrichter, and W. A. Eaton. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.*, 14(1):76–88, 2004.

[100] A. Berezhkovskii and A. Szabo. One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *J. Chem. Phys.*, 122(1):014503, 2005.

[101] R. B. Best and G Hummer. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys. Rev. Lett.*, 96(22):228104, 2006.

[102] S. V. Krivov and M. Karplus. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 105(37):13841–13846, 2008.

[103] H. S. Chung, J. M. Louis, and W. A. Eaton. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. U.S.A.*, 106(29):11837–11844, 2009.

[104] R. B. Best and G. Hummer. Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 107(3):1088–1093, 2010.

[105] S. V. Krivov. Is protein folding sub-diffusive? *PLoS Comput. Biol.*, 6(9):e1000921, 2010.

[106] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[107] R. B. Best and G. Hummer. Diffusion models of protein folding. *Phys. Chem. Chem. Phys.*, 13(38):16902–16911, 2011.

[108] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[109] H. Yu, A. N. Gupta, X. Liu, K. Neupane, A. M. Brigley, I. Sosova, and M. T. Woodside. Energy landscape analysis of native folding of the prion protein yields the diffusion constant, transition path time, and rates. *Proc. Natl. Acad. Sci. U.S.A.*, 109(36):14452–14457, 2012.

[110] H. S. Chung, K. McHale, J. M. Louis, and W. A. Eaton. Single-molecule fluorescence experiments determine protein folding transition path times. *Science*, 335(6071):981–984, 2012.

[111] H. S. Chung and W. A. Eaton. Single-molecule fluorescence probes dynamics of barrier crossing. *Nature*, 502(7473):685–688, 2013.

[112] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.*, 133(45):18413–18419, 2011.

[113] H. Gelman and M. Gruebele. Fast protein folding kinetics. *Q. Rev. Biophys.*, 47(2):95–142, 2014.

[114] T. Schlick, R. Collepardo-Guevara, L. A. Halvorsen, S. Jung, and X. Xiao. Biomolecular modeling and simulation: A field coming of age. *Q. Rev. Biophys.*, 44(2):191–228, 2011.

[115] R. O. Dror, M. Ø. Jensen, D. W. Borhani, and David E. Shaw. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J. Gen. Physiol.*, 135(6):555–562, 2010.

[116] M. Karplus and G. A. Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, 1990.

[117] A. T. Brünger, J. Kuriyan, and M. Karplus. Crystallographic R factor refinement by molecular dynamics. *Science*, 235(4787):458–460, 1987.

[118] A. Vitalis and A. Caflisch. Equilibrium sampling approach to the interpretation of electron density maps. *Structure*, 22(1):156–167, 2014.

[119] M. R. Jensen, R. W. H. Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.*, 23(3):426–435, 2013.

[120] M. R. Jensen, M. Zweckstetter, J. Huang, and M. Blackledge. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev.*, 114(13):6632–6660, 2014.

[121] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, and H. B. Yu. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed.*, 45(25):4064–4092, 2006.

[122] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.*, 9(9):646–652, 2002.

[123] C. Mura and C. E. McAnany. An introduction to biomolecular simulations and docking. *Mol. Simul.*, 40(10–11):732–764, 2014.

[124] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[125] A. D. MacKerell Jr. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.*, 25(13):1584–1604, 2004.

[126] S. Piana, J. L. Klepeis, and D. E. Shaw. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.*, 24:98–105, 2014.

[127] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[128] B. R. Brooks, C. L. Brooks III, A. D. MacKerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[129] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[130] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(16):585–590, 1977.

[131] S. A. Adcock and J. A. McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106(5):1589–1615, 2006.

[132] H. A. Scheraga, M. Khalili, and A. Liwo. Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.*, 58:57–83, 2007.

[133] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.

[134] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[135] C. Hyeon and D. Thirumalai. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.*, 2:487, 2011.

[136] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth. The theory of ultra-coarse-graining. 1. General principles. *J. Chem. Theory Comput.*, 9(5):2466–2480, 2013.

[137] K. Meier, A. Choutko, J. Dolenc, A. P. Eichenberger, S. Riniker, and W. F. van Gunsteren. Multi-resolution simulation of biomolecular systems: A review of methodological issues. *Angew. Chem., Int. Ed.*, 52(10):2820–2834, 2013.

[138] M. Feig and C. L. Brooks III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, 14(2):217–224, 2004.

[139] J. Chen, C. L. Brooks III, and J. Khandogin. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.*, 18(2):140–148, 2008.

[140] J. Kleinjung and F. Fraternali. Design and application of implicit solvent models in biomolecular simulations. *Curr. Opin. Struct. Biol.*, 25:126–134, 2014.

[141] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[142] R. Elber. Long-timescale simulation methods. *Curr. Opin. Struct. Biol.*, 15(2):151–156, 2005.

[143] H. Lei and Y. Duan. Improved sampling methods for molecular simulation. *Curr. Opin. Struct. Biol.*, 17(2):187–191, 2007.

[144] D. M. Zuckerman. Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.*, 40:41–62, 2011.

[145] M. Bacci, A. Vitalis, and A. Caflisch. A molecular simulation protocol to avoid sampling redundancy and discover new states. *Biochim. Biophys. Acta, Gen. Subj.*, 1850(5):889–902, 2014.

[146] M. Shirts and V. S. Pande. Screen savers of the world, unite. *Science*, 290(5498):1903–1904, 2000.

[147] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Proc. 34th Annu. Int. Symp. Comput. Archit.*, pages 1–12, 2007.

[148] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[149] M. J. Harvey, G. Giupponi, and G. De Fabritiis. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.*, 5(6):1632–1639, 2009.

[150] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.*, 8(5):1542–1555, 2012.

[151] M. J. Harvey and G. De Fabritiis. A survey of computational molecular science using graphics processing units. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(5):734–742, 2012.

[152] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics: Model problems and algorithms. *Nonlinearity*, 17(6):R55, 2004.

[153] G. S. Buchner, R. D. Murphy, N.-V. Buchete, and J. Kubelka. Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory. *Biochim. Biophys. Acta, Proteins Proteomics*, 1814(8):1001–1020, 2011.

[154] M. A. Rohrdanz, W. Zheng, and C. Clementi. Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu. Rev. Phys. Chem.*, 64:295–316, 2013.

[155] I. Jolliffe. *Principal component analysis*. Springer, New York, 2nd edition, 2002.

[156] S. Hayward and N. Go. Collective variable description of native protein dynamics. *Annu. Rev. Phys. Chem.*, 46:223–250, 1995.

[157] T. F. Cox and M. A. A. Cox. *Multidimensional scaling*. Chapman & Hall / CRC, Boca Raton, 2000.

[158] M. Levitt. Molecular dynamics of native protein: II. Analysis and nature of motion. *J. Mol. Biol.*, 168(3):621–657, 1983.

[159] P. S. Penev and J. J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Comp. Neural*, 7(3):477–500, 1996.

[160] Z. Zhang and W. Wriggers. Local feature analysis: A statistical theory for reproducible essential dynamics of large macromolecules. *Proteins: Struct., Funct., Bioinf.*, 64(2):391–403, 2006.

[161] Y. Xue, P. J. Ludovice, M. A. Grover, L. V. Nedialkova, C. J. Dsilva, and I. G. Kevrekidis. State reduction in molecular simulations. *Comput. Chem. Eng.*, 51:102–110, 2013.

[162] O. F. Lange and H. Grubmüller. Full correlation analysis of conformational protein dynamics. *Proteins: Struct., Funct., Bioinf.*, 70(4):1294–1312, 2008.

[163] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72(23):3634–3637, 1994.

[164] Y. Naritomi and S. Fuchigami. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.*, 134(6):065101, 2011.

[165] C. R. Schwantes and V. S. Pande. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.*, 9(4):2000–2009, 2013.

[166] G. Pérez-Hernández, F. Paul, T. Giorgino, G. de Fabritiis, and F. Noé. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.*, 139(1):015102, 2013.

[167] Y. Naritomi and S. Fuchigami. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.*, 139(21):215102, 2013.

[168] C. R. Schwantes and V. S. Pande. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.*, 11(2):600–608, 2015.

[169] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[170] U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

[171] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.*, 107(31):13597–13602, 2010.

[172] A. L. Ferguson, S. Zhang, I. Dikiy, A. Z. Panagiotopoulos, P. G. Debenedetti, and A. J. Link. An experimental and computational investigation of spontaneous lasso formation in microcin J25. *Biophys. J.*, 99(9):3056–3065, 2010.

[173] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134(12):124116, 2011.

[174] W. Zheng, B. Qi, M. A. Rohrdanz, A. Caflisch, A. R. Dinner, and C. Clementi. Delineation of folding pathways of a β-sheet miniprotein. *J. Phys. Chem. B*, 115(44):13065–13074, 2011.

[175] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem. Phys. Lett.*, 509(1–3):1–11, 2011.

[176] H. J. C. Berendsen and S. Hayward. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.*, 10(2):165–169, 2000.

[177] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21(1):113–127, 2006.

[178] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.*, 7(2):842–864, 2008.

[179] M. Duan, J. Fan, M. Li, L. Han, and S. Huo. Evaluation of dimensionality-reduction methods from peptide folding-unfolding simulations. *J. Chem. Theory Comput.*, 9(5):2490–2497, 2013.

[180] A. Vitalis and A. Caflisch. Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theory Comput.*, 8(3):1108–1120, 2012.

[181] A. Caflisch. Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.*, 16(1):71–78, 2006.

[182] L. S. Ahlstrom, J. L. Baker, K. Ehrlich, Z. T. Campbell, S. Patel, I. I. Vorontsov, F. Tama, and O. Miyashita. Network visualization of conformational sampling during molecular dynamics simulation. *J. Mol. Graphics Modell.*, 46:140–149, 2013.

[183] S. V. Krivov and M. Karplus. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B*, 110(25):12689–12698, 2006.

[184] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B*, 112(29):8701–8714, 2008.

[185] O. M. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106(4):1495–1517, 1997.

[186] S. V. Krivov and M. Karplus. Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys.*, 117(23):10894–10903, 2002.

[187] L. C. Smeeton, M. T. Oakley, and R. L. Johnston. Visualizing energy landscapes with metric disconnectivity graphs. *J. Comput. Chem.*, 35(20):1481–1490, 2014.

[188] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 104(6):1817–1822, 2007.

[189] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Struct., Funct., Bioinf.*, 70(4):1185–1195, 2008.

[190] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126(15):155101, 2007.

[191] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, 18(2):154–162, 2008.

[192] W. E and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.*, 61:391–420, 2010.

[193] R. T. McGibbon, C. R. Schwantes, and V. S. Pande. Statistical model selection for Markov models of biomolecular dynamics. *J. Phys. Chem. B*, 118(24):6475–6481, 2014.

[194] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.*, 139(18):184114, 2013.

[195] R. T. McGibbon, B. Ramsundar, M. M. Sultan, G. Kiss, and V. S. Pande. Understanding protein dynamics with L_1-regularized reversible hidden Markov models. *Proc. 31st Int. Conf. Mach. Learn.*, pages 1197–1205, 2014.

[196] G. R. Bowman, L. Meng, and X. Huang. Quantitative comparison of alternative methods for coarse-graining biological networks. *J. Chem. Phys.*, 139(12):121905, 2013.

[197] F. Noé and F. Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.*, 11(2):635–655, 2013.

[198] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé. Variational approach to molecular kinetics. *J. Chem. Theory Comput.*, 10(4):1739–1752, 2014.

[199] W. Li and A. Ma. Recent developments in methods for identifying reaction coordinates. *Mol. Simul.*, 40(10-11):784–793, 2014.

[200] B. Qi, S. Muff, A. Caflisch, and A. Dinner. Extracting physically intuitive reaction coordinates from transition networks of a β-sheet miniprotein. *J. Phys. Chem. B*, 114(20):6979–6989, 2010.

[201] S. V. Krivov. The free energy landscape analysis of protein (FIP35) folding dynamics. *J. Phys. Chem. B*, 115(42):12315–12324, 2011.

[202] P. V. Banushkina and S. V. Krivov. High-resolution free-energy landscape analysis of α-helical protein folding: HP35 and its double mutant. *J. Chem. Theory Comput.*, 9(12):5257–5266, 2013.

[203] A. M. Berezhkovskii and A. Szabo. Diffusion along the splitting/commitment probability reaction coordinate. *J. Phys. Chem. B*, 117(42):13115–13119, 2013.

[204] R. Hegger and G. Stock. Multidimensional Langevin modeling of biomolecular dynamics. *J. Chem. Phys.*, 130(3):034106, 2009.

[205] N. Schaudinnus, A. J. Rzepiela, R. Hegger, and G. Stock. Data driven Langevin modeling of biomolecular dynamics. *J. Chem. Phys.*, 138(20):204106, 2013.

[206] A. Ljubetič, I. Urbančič, and J. Štrancar. Recovering position-dependent diffusion from biased molecular dynamics simulations. *J. Chem. Phys.*, 140(8):084109, 2014.

[207] Z. Lai, K Zhang, and J. Wang. Exploring multi-dimensional coordinate-dependent diffusion dynamics on the energy landscape of protein conformation change. *Phys. Chem. Chem. Phys.*, 16(14):6486–6495, 2014.

[208] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.*, 101(41):14766–14770, 2004.

[209] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. *SIGMOD '99 Proc. 1999 ACM SIGMOD Int. Conf. Manage. Data*, pages 49–60, 1999.

[210] K. Müller and L. D. Brown. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theor. Chim. Acta*, 53(1):75–93, 1979.

[211] K. Müller. Reaction paths on multidimensional energy hypersurfaces. *Angew. Chem., Int. Ed.*, 19(1):1–13, 1980.

[212] W. E, W. Ren, and E. Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.*, 126(16):164103, 2007.

[213] T. Sauer. Computational solution of stochastic differential equations. *Wiley Interdiscip. Rev. Comput. Stat.*, 5(5):362–371, 2013.

[214] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[215] A. Wlodawer, J. Walter, R. Huber, and L. Sjölin. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Biol.*, 180(2):301–329, 1984.

[216] M. J. Grey, C. Wang, and A. G. Palmer III. Disulfide bond isomerization in basic pancreatic trypsin inhibitor: Multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J. Am. Chem. Soc.*, 125(47):14324–14335, 2003.

[217] Y. Xue, J. M. Ward, T. Yuwen, I. S. Podkorytov, and N. R. Skrynnikov. Microsecond time-scale conformational exchange in proteins: Using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. *J. Am. Chem. Soc.*, 134(5):2555–2562, 2012.

[218] D. Huang and A. Caflisch. The evolutionary conserved Tyr169 stabilizes the β2-α2 loop of the prion protein. *J. Am. Chem. Soc.*, 137(8):2948–2957, 2015.

[219] N. Blöchliger, M. Xu, and A. Caflisch. Peptide binding to a PDZ domain by electrostatic steering via non-native salt bridges. *Biophys. J.*, 108(9):2362–2370, 2015.

[220] S. Wallin, J. Farwer, and U. Bastolla. Testing similarity measures with continuous and discrete protein models. *Proteins: Struct., Funct., Bioinf.*, 50(1):144–157, 2003.

[221] T. Zhou and A. Caflisch. Distribution of reciprocal of interatomic distances: A fast structural metric. *J. Chem. Theory Comput.*, 8(8):2930–2937, 2012.

[222] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data*, 3(1):1, 2009.

[223] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos. Locally adaptive metrics for clustering high dimensional data. *Data Min. Knowl. Disc.*, 14(1):63–97, 2007.

[224] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. *Proc. 2004 SIAM Int. Conf. Data Mining*, pages 517–521, 2004.

[225] P. Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87, 2012.

[226] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[227] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 32(5):922–923, 1976.

[228] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[229] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U.S.A.*, 103(26):9885–9890, 2006.

[230] B. Hashemian, D. Millán, and M. Arroyo. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *J. Chem. Phys.*, 139(21):214101, 2013.

[231] M. Ceriotti, G. A. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.*, 108(32):13023–13028, 2011.

[232] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998.

[233] E. P. Xing, M. I. Jordan, S. Russell, and Andrew Y. Ng. Distance metric learning with application to clustering with side-information. *Adv. Neural Inf. Process. Syst.*, 15:521–528, 2002.

[234] R. T. McGibbon and V. S. Pande. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theory Comput.*, 9(7):2900–2906, 2013.

[235] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. *IEEE T. Pattern Anal. Mach. Intell.*, 24(9):1281–1285, 2002.

[236] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *J. Roy. Stat. Soc. B*, 66(4):815–849, 2004.

[237] L. Jing, M. K. Ng, and J. Z. Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE T. Knowl. Data En.*, 19(8):1026–1041, 2007.

[238] Y. Li, M. Dong, and J. Hua. Localized feature selection for clustering. *Pattern Recogn. Lett.*, 29(1):10–18, 2008.

[239] A. Singer and R. R. Coifman. Non-linear independent component analysis with diffusion maps. *Appl. Comput. Harmon. Anal.*, 25(2):226–239, 2008.

[240] A. Singer, R. Erban, I. G. Kevrekidis, and R. R. Coifman. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.*, 106(38):16090–16095, 2009.

[241] C. J. Dsilva, R. Talmon, N. Rabin, R. R. Coifman, and I. G. Kevrekidis. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *J. Chem. Phys.*, 139(18):184109, 2013.

Chapter 2

A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems

Blöchliger, N., Vitalis, A. and Caflisch, A. *Computer Physics Communications*, 184(11): 2446–2453, 2013

Computer Physics Communications 184 (2013) 2446–2453

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems

Nicolas Blöchliger, Andreas Vitalis *, Amedeo Caflisch

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
7 June 2013
Accepted 12 June 2013
Available online 25 June 2013

Keywords:
Complex system
Trajectory analysis
Scalable algorithm
Minimum spanning tree
Free energy basins

ABSTRACT

Advances in IT infrastructure have enabled the generation and storage of very large data sets describing complex systems continuously in time. These can derive from both simulations and measurements. Analysis of such data requires the availability of scalable algorithms. In this contribution, we propose a scalable algorithm that partitions instantaneous observations (snapshots) of a complex system into kinetically distinct sets (termed basins). To do so, we use a combination of ordering snapshots employing the method's only essential parameter, *i.e.*, a definition of pairwise distance, and annotating the resultant sequence, the so-called progress index, in different ways. Specifically, we propose a combination of cut-based and structural annotations with the former responsible for the kinetic grouping and the latter for diagnostics and interpretation. The method is applied to an illustrative test case, and the scaling of an approximate version is demonstrated to be $\mathcal{O}(N \log N)$ with N being the number of snapshots. Two real-world data sets from river hydrology measurements and protein folding simulations are then used to highlight the utility of the method in finding basins for complex systems. Both limitations and benefits of the approach are discussed along with routes for future research.

1. Introduction

With present day computing resources, large-scale temporal simulations of complex systems can be performed routinely, and time-resolved, experimental data in many dimensions are collected and stored. In both cases, the resultant, very large amounts of data require dedicated, scalable protocols to handle access and analysis [1–3]. Examples can be found in fields such as protein science [4,5], astronomy [6], cell biology [7], or climatology [8] to name just a few.

For a complex system evolving in time, data are present in the form of sequences of instantaneous snapshots (microstates in the language of statistical mechanics) of this complex system, and such a sequence will be referred to as a trajectory throughout. Depending on whether data are synthetic or real, the implied projection of the system to obtain a snapshot may differ, and this may limit spatial resolution. Temporal resolution is limited directly by the instruments or numerical schemes if storage space is not a concern. Even though continuous evolution need not be observed explicitly as a function of time, we will restrict our terminology to this case. Routine analyses of trajectory data may involve computing average properties and their estimated distribution functions in $\mathcal{O}(N)$ time, where N is the number of snapshots. Distribution functions offer hints toward the diversity of states visited by the complex system and their relative weights. Time-resolved analyses provide insight regarding state connectivity and transition rates. Projection onto low-dimensional properties is necessary to render such analyses statistically meaningful and visualizable by conventional means.

If we assume that snapshots follow a well-defined distribution (such as the Boltzmann distribution for particles in the classical limit), these analyses look for spatial domains that are highly populated under the given conditions, *i.e.*, those for which a finite sample yields higher-than-average densities of microstates, preferably through recurrence [9]. Here, recurrence refers to the trajectory's property of entering and exiting subdomains within high density regions several times. The motivation behind this is twofold: (1) characterization of the complex system and communication of results in terms fit for human consumption [10]; (2) derivation of simplified models that provide a meaningful representation of the complex system [11,12]. Such models can preserve coarse-grained dynamical and static properties of the system and enable predictions to be made over vastly extended temporal or spatial domains.

When analyzing trajectories in projected spaces, high density regions are prone to overlap, and plots rarely resolve all of them [13]. This overlap phenomenon can lead to incorrect conclusions regarding the diversity and connectivity of coarse states. Consequently, affordable protocols that require little knowledge of the system *a priori* and that decrease the likelihood of such overlap are of interest. Techniques such as principal component analysis,

* Corresponding author. Tel.: +41 446355597; fax: +41 446356862.
E-mail addresses: n.bloechliger@bioc.uzh.ch (N. Blöchliger), a.vitalis@bioc.uzh.ch (A. Vitalis), caflisch@bioc.uzh.ch (A. Caflisch).

0010-4655/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2013.06.009

spectral clustering [14] and the related diffusion maps [15], locally linear embeddings [16], cut-based free energy profiles [17], kinetic groupings based on networks [18–21], which are specific cases of community detection algorithms in graphs [22], *etc.* are all in use, but many of them scale superlinearly with N.

Data clustering [23] offers a simple route to the identification of high density domains. Clusters are defined as groups of mutually similar snapshots. Similarity is assessed by a criterion of distance generally requiring an *ad hoc* selection of both a subset of features [24] and a functional form. However, a grouping meant to describe an evolving system should also encode dynamic proximity [25], *i.e.*, given a time resolution, which snapshots constitute a kinetically distinct state? If the system is of atomic scale and at equilibrium, this question aims to identify free energy basins and barriers in a generally high-dimensional phase space [26,27]. Positional coordinates of atoms are often used exclusively given that momenta can likely be ignored out on account of their much shorter autocorrelation times. We note that the language and concepts of statistical physics have proven useful in the analysis of nonphysical systems as well [28], *i.e.*, our adaptation of this language does not imply a restricted domain of application.

In this contribution, we present an algorithm that operates directly on a trajectory. With just the definition of a pairwise distance between snapshots, we are able to generate a one-dimensional plot that allows the identification of states in a joint geometric and kinetic sense, which we will refer to as basins. With standard metrics derived from microstate representations (such as interatomic distances in a flexible molecule), the method relies on the continuity of geometric representations in time. This implies that it may fail for certain classes of discrete systems. The main benefits of our algorithm are that it does not rely on any parameters *per se*, that it is very likely to resolve all basins, and that with the help of reasonable approximations to the exact procedure, the total running time approaches $\mathcal{O}(N \log N)$. The combination of these points is worth emphasizing, since we believe that they constitute a desirable and unique fingerprint of our approach.

The rest of this manuscript is structured as follows. First, we present the key ideas behind the procedure (Section 2.1) and illustrate its utility with a suitable model system (2.2). Next, we describe a computationally efficient and robust approximation to the exact procedure. The scaling of computational cost with data set size and dimensionality is tested explicitly (2.3). This is followed by applying the method to two complex real-world data sets, the first from hydrology (3.1) and the second from protein folding (3.2). We conclude by discussing the advantages and possible problems in comparison with related approaches (4).

2. Methods and proof of concept

2.1. The exact algorithm

Let $T = \{t_1, \ldots, t_N\}$ be a set (trajectory) of N unique snapshots, which usually are representations of the system in $\mathbb{R}^D$, which is the chosen subspace of the original system representation with $D \leq D_{system}$. We use any pairwise distance $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ to measure the similarity between two snapshots. This need not be a purely coordinate-dependent function. Below it will prove beneficial for d to be a metric yielding a continuous number space with all $\mathcal{O}(N^2)$ values of d being unique.

We can now define the following iterative procedure. Choose a starting snapshot $s_1 \in T$ and create the set $S_1 = \{s_1\}$. Initialize the cut function, $c : \{1, \ldots, N\} \to \mathbb{N}$, to 2. Then, for $i = 1, \ldots, N-1$ do the following:

1. Define s_{i+1} as the snapshot in $T \backslash S_i$ realizing the minimum of $d(\cdot, S_i) = \min_{j=1,\ldots,i} d(\cdot, s_j)$.

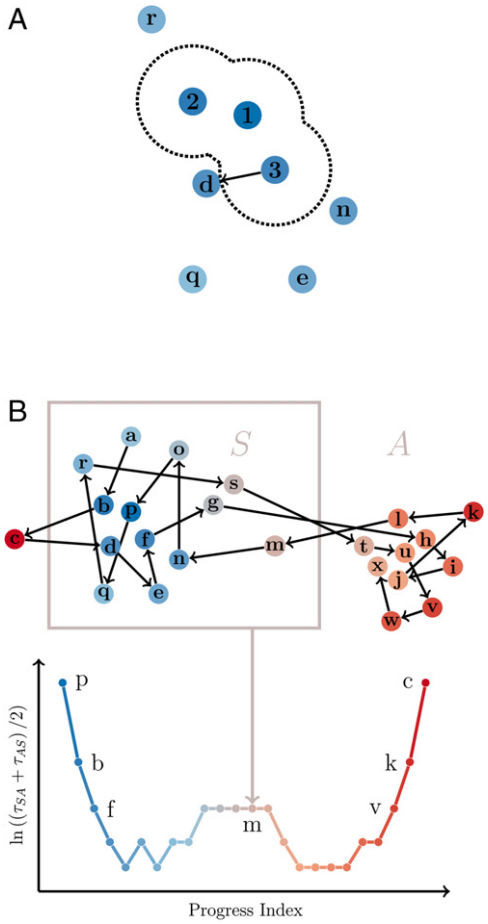


Fig. 1. Schematic highlighting the fundamental components of the algorithm. **A**. A set of points in two dimensions is shown as circles. See 2.1 for details. **B**. The points in **A** are shown as a subset of a larger data set. Arrows and letters indicate progression in time. The color scheme follows the order in which points are added when starting with point **p**, *i.e.*, colors trace the progress index itself. The schematic on the bottom shows values for the inverse logarithm of c at each value of the progress index. An example point and the cut to obtain the respective partitions S_i and A_i are highlighted. Point **c** illustrates an outlier, which are prone to be added last to S. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Let $S_{i+1} = S_i \cup \{s_{i+1}\}$.
3. Define $c(i+1) = \sum_{j=1}^{N-1} \zeta_{S_{i+1}}(t_j, t_{j+1})$.

Here, the function ζ is defined as

$$\zeta_X(t,u) = \begin{cases} 0 & \text{if neither or both } t \text{ and } u \text{ are part of set } X \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

The exact progress index of T starting with s_1 is defined as the sequence $S(T, s_1) = (s_1, \ldots, s_N)$. Each entry i is associated with a value for the cut function, $c(i)$. In words, given a starting snapshot, the algorithm finds a unique ordering of the snapshots, and annotates it with the number of transitions between the two partitions defined by all the snapshots that are currently part of the set (S_i) and those that are not yet part of the set ($A_i = T \backslash S_i$). The cut function c is related to the mean first passage time in the implied two-state Markov model via

$$\tau_{\mathrm{MFP}}(A_i \to S_i) + \tau_{\mathrm{MFP}}(S_i \to A_i) = 2N/c(i). \quad (2)$$

We use τ_{AS} as shorthand notation for $\tau_{\mathrm{MFP}}(A_i \to S_i)$ throughout. In Fig. 1(A), we show an illustration of a trajectory in 2D space with the current set of snapshots 1–3. The order of adding further snapshots would then be **d**, **n**, **r**, **e**, and **q** based on the mutual distance relations. There are no free parameters beyond having to

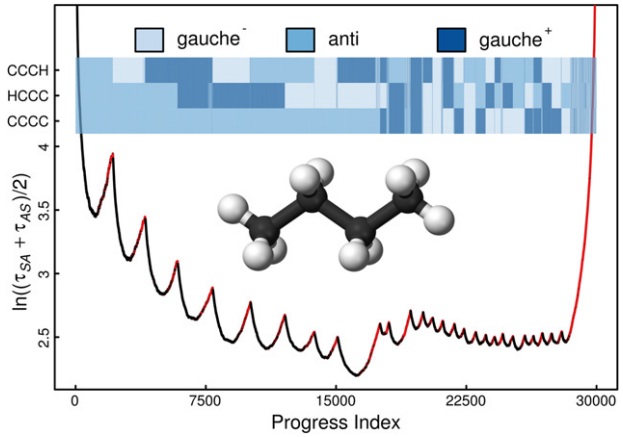


Fig. 2. Illustration of the approach using *n*-butane. The 27 basins of the system are all clearly resolved. Amongst those basins with the CCCC dihedral angle in *anti*, adjacent basins involve the rotation of only one of the methyl groups. This is fortuitous but signifies that the following basin in terms of the progress index is chosen on account of the sampling density in transition regions to any of the preceding ones. This density is higher for transitions involving only a single rotation. Points plotted in red correspond to snapshots that are classified as eclipsed according to the binning strategy described in 2.2 and are found preferentially toward the right half of basins and at the largest values of the progress index in general. The color annotation uses a simplified binning into 120° bins and does not display eclipsed microstates. The implied unit of time on the *y*-axis is a single snapshot, *i.e.*, 250 fs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

define distance relations, and for this purpose we have chosen the canonical tool, *i.e.*, a metric. In Fig. 1(B), the same set of points is shown as part of a longer trajectory. Here, letters indicate temporal order (**a**–**x**), whereas coloring tracks the progress index (blue–red) when using **p** as the starting snapshot. The cut function, *i.e.*, the number of transitions between S_i and $T\backslash S_i$, is illustrated in the lower half of the plot. The logarithm of the inverse of $c(i)$ produces small values if there are many transitions and peaks if there are few transitions. The latter is highlighted in Fig. 1(B) for set S with **m** being the snapshot having been added last. Fig. 1(B) illustrates the hypothesis that maxima in the logarithm of Eq. (2) will correspond to kinetic barriers separating basins to the left from those to the right. Consequently, the cut function should qualitatively encode dynamic properties of the system.

We note that the algorithm has two distinct parts: the progress index generation and the annotation function, here the cut function c. Both components can be treated and modified independently. A determination of the exact progress index is related to finding the minimum spanning tree (MST) of a complete graph with N vertices corresponding to all the t_i and edges with weights given via $d(t_i, t_j)$. The implementation we use scales with an overall complexity near $\mathcal{O}(N^2)$ and is described briefly in the Supplementary Information (SI), S.1.1 (see the Appendix). The exact progress index of T is unique if all possible $d(t_i, t_j)$ are distinct, and a unique progress index does not depend on the order the snapshots appear in T, *i.e.*, it does not contain any kinetic information. By construction, it is not possible for geometrically distinct basins to overlap provided that the sampling is good enough. Moreover, it is worth noting that the progress index does not imply that a given basin is closest kinetically to the one immediately to the left, but rather to any basin to the left.

2.2. Illustration with labeled n-butane

Let the linear alkane *n*-butane be described by three dihedral angles specifying rotations around all three carbon–carbon bonds (see Fig. 2). We assume atoms to be labeled such that the degeneracy of states can be resolved. In our chosen description, each dihedral angle has three distinct potential energy minima at 180°, 60°, and −60° corresponding to *anti*, *gauche*$^+$, and *gauche*$^-$ conformations. The potential has threefold symmetry for the methyl groups but favors *anti* for the central dihedral angle. It is expected that the system has access to $3^3 = 27$ coarse, metastable states. This is a good example for the algorithm presented in 2.1 since the low dimensionality and good knowledge of the system allow us to characterize basins and transitions independently.

Using stochastic dynamics simulations (see SI, S.1.4.1), we generated a classical trajectory of 30 000 snapshots under conditions such that recurrent sampling of all 27 basins is observed. Fig. 2 shows a plot generated by the algorithm described in 2.1 based on a trajectory with a time resolution of 250 fs and using a distance function defined on the three dihedral angles [29]. Clearly, we can resolve all basins, which is in contrast to cut-based free energy profiles used in prior work [29]. To confirm that the indicated basins do indeed correspond to the 27 expected ones, a color map representing an independent annotation based on binning the three degrees of freedom separately is shown. This correspondence is also established in Fig. S.1 with the help of box plots. Both figures reveal an asymmetry for snapshots within basins: points in highest density regions appear toward the left, and points in lower density ("fringe") regions appear toward the right. The latter correspond to eclipsed states, which have maximal enthalpy for this system. The asymmetry within each basin is a natural consequence of the way the progress index is constructed and annotated.

Further exploration of this system is meant to analyze two critical issues. First, what is the impact of the trajectory's time resolution? Second, can a connection between the results in Fig. 2 and an independent analysis of the thermodynamics and kinetics of this system be established?

We expect the progress index annotated with c as in Fig. 2 to successively lose its pertinent features if time resolution becomes so coarse that the various basin-to-basin transitions can no longer be resolved. We note that such a trajectory will eventually look random, which implies that the cut function just reports on the relative sizes of the two partitions, and not on (time-)local groupings of snapshots. This is indeed the case as shown in Fig. 3. For a resolution beyond 6 ps, the profile relaxes to a smooth, parabolic shape, which can be rationalized based on combinatorial arguments. We plot as a dashed line in Fig. 3 the analytically derived prior expected for a completely random trajectory (see SI, S.1.2). The result in Fig. 3 is obtained despite the fact that the progress index still orders the snapshots in fundamentally the same way as at finer time resolution. To make this point, a color map analogous to Fig. 2 is shown in Fig. 3 for the progress index derived from the 31.25 ps case. Therefore, the lack of features in Fig. 3 is not a result of overlap in the way one would encounter it in histogram- [17,30] or cut-based methods [29]. This is a significant advantage of our approach.

To perform an independent analysis of thermodynamics and kinetics, we constructed a set of macrostates by creating a 3D histogram with cubic bins of side length 60°. Bins are called eclipsed unless all three dimensions are centered at one of the three potential energy minima. Thus, 3^3 out of 6^3 macrostates are not eclipsed, and those correspond to the 27 basins. The resultant sequence of macrostates can be used to infer the transition matrix of an underlying Markov state model (MSM). From the MSM, pairwise τ_{MFP} values can be computed. If we now consider the progress index, at each point, we have a given MSM state annotation of the points immediately to the left (smaller values of the progress index) and to the right (larger values of the progress index). We may then infer the dominantly populated macrostate to either side via maximum likelihood guesses. With the knowledge of those two guesses, L and R, for each point of the progress index, we can plot the sum $\tau_{\mathrm{MFP}}(L \to R) + \tau_{\mathrm{MFP}}(R \to L)$. If $L \equiv R$, the result is directly proportional to the inverse of the probability of $L \equiv R$. Conversely,

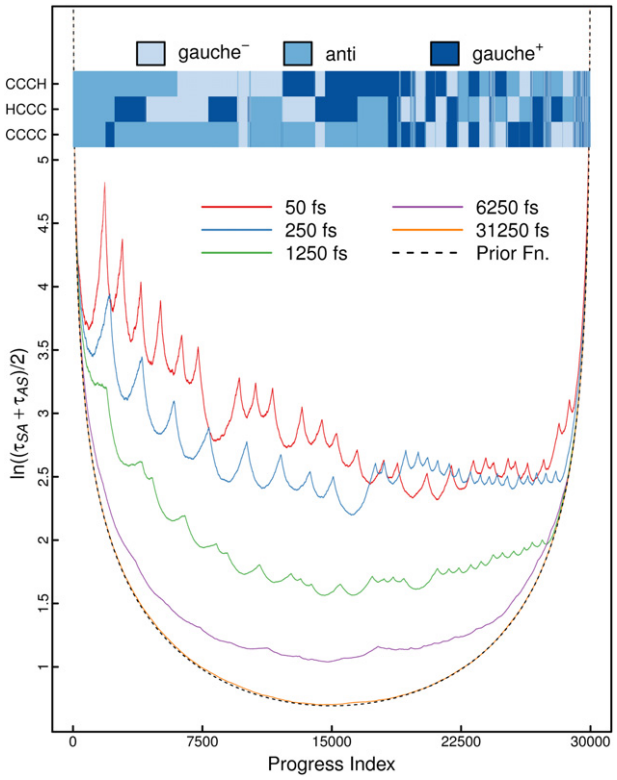


Fig. 3. Dependence of progress index and annotation function on temporal resolution. Data comparable to Fig. 2 are shown for decreasing temporal resolution. Features are successively lost, and at 31.25 ps the annotation becomes indistinguishable from that expected for a completely random trajectory (prior function). For the cases of 1.25 and 6.25 ps, it is apparent that the strong inherent curvature of function c impedes the identification of basins if they are small and/or temporal resolution is poor. For each curve the implied unit of time on the y-axis is a single snapshot of the respective trajectory, *i.e.*, the saving frequency or temporal resolution itself. As in Fig. 2, a color annotation is shown, here for the 31.25 ps case. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

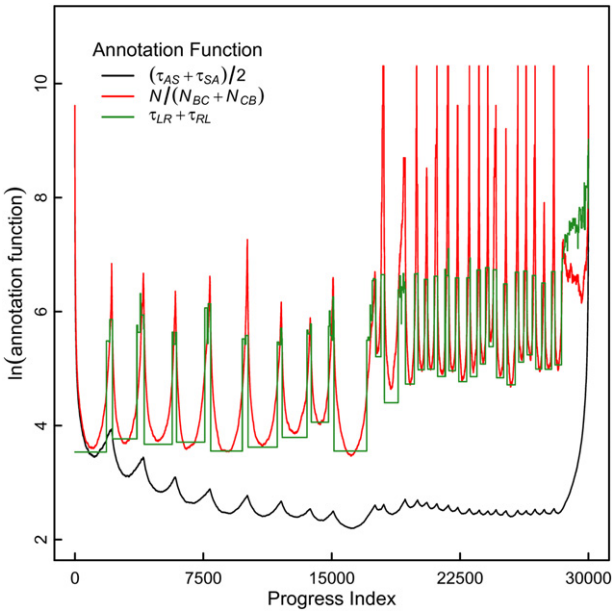


Fig. 4. Quantitative kinetic and thermodynamic interpretation of two annotation functions. The standard annotation function via Eq. (1) is reproduced identically to Fig. 2 (black). The localized annotation function defined in Eq. (3) is shown in red. Because basins have two standard sizes (assumed to be 1600 snapshots if the central torsion is in *anti* and 400 snapshots otherwise), we generated data with $n_l(i)$ set to fixed values of either 1600 or 400 snapshots. For the curve shown in the plot, values were simply interpolated to convert from the case with $n_l(i) = 1600$ to the case with $n_l(i) = 400$ over values of the progress index of 17 300–17 700. Lastly, the green curve shows results from an underlying MSM as described in 2.2. The width for constructing the maximum likelihood guess of assigning basins L and R was 100 snapshots throughout. The implied unit of time on the y-axis is a single snapshot, *i.e.*, 250 fs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

if the point is in a barrier region, $L \neq R$ and the result measures the kinetic proximity of two neighboring macrostates. These data are shown as the green curve in Fig. 4. Comparison with the original profile shows that there is no quantitative relationship between the two plots. It is therefore impossible to obtain quantitative thermodynamic or kinetic information from c. This is expected because the cut function measures kinetics in a crude two-state assumption (A and S above) and not between individual basins.

Are there alternative annotation functions to consider? Here, we define a 'localized' cut function as follows:

$$l(i) = \sum_{j=1}^{N-1} \zeta_{B_i(n_l(i))}(t_j, t_{j+1}) \zeta_{C_i(n_l(i))}(t_j, t_{j+1}). \tag{3}$$

In Eq. (3), partition $B_i(n_l(i))$ is defined as $S_{i-1} \backslash S_{i-1-n_l(i)}$, and partition $C_i(n_l(i))$ is defined as $S_{i-1+n_l(i)} \backslash S_{i-1}$. This corresponds to a restriction of the cut function to contributions from points in the trajectory that are near in the progress index, and function l is expected to offer better resolution than c for reasonable choices of $n_l(i)$. A progress index annotated with l is shown in Fig. 4 as well. Due to the peculiar nature of the system, the parameter $n_l(i)$ in Eq. (3) is chosen in accordance with average basin sizes (see the caption to Fig. 4). There is very good correspondence between these results and the thermodynamic information inferred from the MSM. However, Fig. 4 shows that peak heights are not correlated beyond both sets appearing to populate two dominant ranges of values. Quantitative correspondence is unlikely because the cut function defined by Eq. (3) is not equivalent to well-defined kinetic information within the underlying three-state MSM. However, function l does appear to be able to provide higher resolution when it comes to identifying basins. This is highlighted by comparison of Figs. 3 and S.2 for the 1.25 ps case, which reveals that the inherent curvature of function c may limit basin delineation before the time resolution approaches characteristic transition times of the system. If meaningful values for $n_l(i)$ can be found, annotation with l is likely to provide more information.

2.3. An approximate algorithm operating in near-linear time

Because the exact algorithm as described in 2.1 and expanded upon in the SI, S.1.1, requires approximately $\mathcal{O}(N^2)$ time, it is impractical for large data sets. In this section, we outline conceptually the implementation of an approximate algorithm that operates in $\mathcal{O}(N \log N)$ time. A detailed description is found in the SI, S.1.3.

Briefly, a spanning tree is constructed with Borůvka's algorithm [31], which works by successively merging subtrees using nearest neighbors. However, instead of considering rigorous nearest neighbors for each subtree, we instead consider a set of nearby snapshots, which is extracted from preorganizing the data via hierarchical clustering [29]. A hierarchical grouping means that snapshots are partitioned into groups of similar objects (clusters) for a range of resolutions. The set of nearby snapshots is then constituted from the union of all clusters that the subtree spans, and which are not yet part of the subtree. This is done for the finest

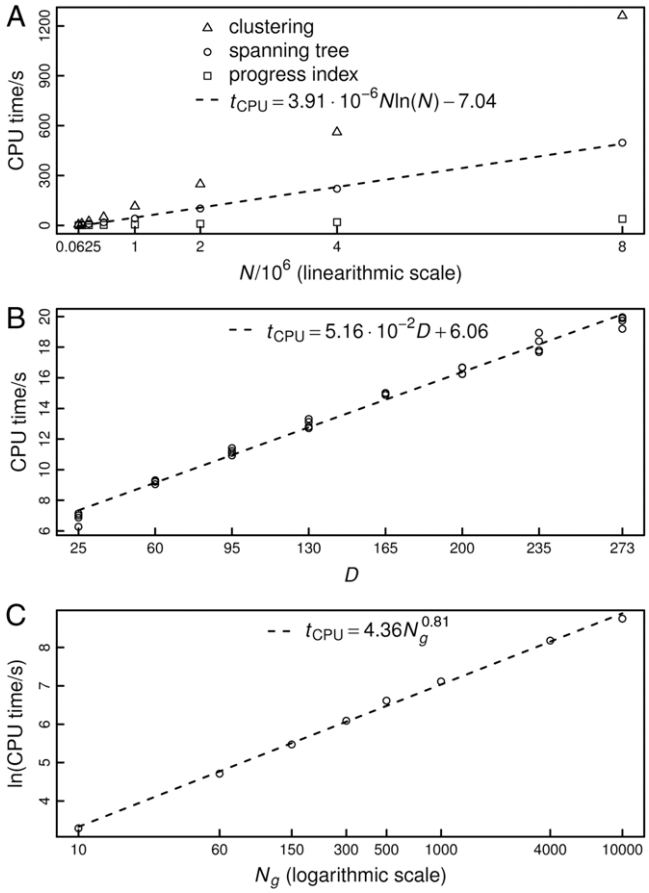


Fig. 5. Runtime analysis for the approximate version of the proposed algorithm. **A**. The cost for computing the SST is shown as a function of $N \log N$, *i.e.*, the expected complexity. We also show a linear fit and the costs for the tree-based clustering and generation of the progress index from the SST. An apparent scaling exponent from a double logarithmic plot of cost *vs.* N (not shown) is 1.15, close to the expected value of 1.08 for this range of values for N. Data are for the case where the number of clusters at the leaf level of the hierarchical clustering grows linearly with N, *i.e.*, the average cluster size is roughly constant (see S.1.3). **B**. Computational cost of SST construction as a function of dimensionality. D was adjusted as described in SI, S.1.4.2. As expected, cost is linear in D, and four independent trials yield similar answers. **C**. Computational cost of SST construction as a function of the number of guesses, N_g. Because N_g will eventually exceed the size of the restricted search space, it is expected that cost scales sublinearly with N_g. This is confirmed by the double logarithmic plot.

resolution level still yielding a nonempty set. If the set is larger than a parameter, N_g, N_g guesses are taken instead of searching the entire set. These two approximations limiting the search space mean that for a given subtree the number of candidate edges is within a constant upper bound, which gives the desired complexity of $\mathcal{O}(N \log N)$. The output is a "short" spanning tree (SST) used for the generation of progress index and annotation functions analogously to the MST in the exact case. Qualitative neighbor relations are expected to be preserved in the SST with the approximations primarily leading to randomization within basins.

The scaling with data set size (see SI, S.1.4.2) is demonstrated in Fig. 5(A) for a fixed value for N_g of 20. Clearly, a plot of computational cost *vs.* $N \log N$ is roughly linear. As can be seen, the cost for the construction of the SST along with the generation of data pairs for progress index and annotation function is less than that of the tree-based clustering. Fig. 5(A) implies that we can identify basins in a data set composed of 8×10^6 snapshots with a dimensionality of $D = 273$ in less than an hour on a single core of a modern desktop machine. Fig. 5(B) shows the dependence of computational cost on D. This is expected to be linear, since the dimensionality of representation only matters for computations of distance, the total number of which is roughly constant. This expectation is confirmed by Fig. 5(B).

3. Results

3.1. Hydrology data for rivers near Portland, Oregon

While *n*-butane is a perfect example for the algorithm, real world data sets may not be, especially if they describe the evolution of a complex system that is not fundamentally stochastic in nature. We constructed an example from hydrology parameters measured at various river sites near Portland, Oregon, USA, over a period of about 5 years. Measured quantities include pH, conductance, discharge (volume flux), temperature, and oxygen content. River parameters are expected to be influenced by ambient weather, specific climatic events such as snowmelt, and local geography. Seasonal patterns generate data sets that are likely to show recurrence (similar seasons in subsequent years give rise to similar river conditions), but that are not random. These data are challenging for the following reasons:

1. Measurements are performed with low accuracy and may contain outliers caused by malfunctioning devices or short-term, local contaminations.
2. Subtle trends observed over multiple years may render conditions locally more similar than compared to analogous times in other years, and recurrence of conditions is weak overall due to the (small) number of years in the data set. This challenges the annotation function that relies on good mixing within a basin.
3. Most measured parameters produce uninformative histograms on their own. In conjunction with the first point, this challenges the geometrical separability of these data, *i.e.*, the pairwise distance spectrum is expected to be relatively featureless (see Fig. S.3).

We note that the data set is small enough ($N = 87\,840$ and $D = 15$) that we can use the exact algorithm. Fig. 6 plots the progress index annotated with both c and l, and the kinetic annotation confirms the challenging nature of these data. Profiles are sparse in well-resolved features and allow the identification of two larger basins with unclear size along with a number of smaller basins, *e.g.*, for values of the progress index around $1.6 \cdot 10^4$ or $8 \cdot 10^4$. The color annotation of the input data supports this interpretation. These data were normalized, centered, and subjected to noise before being fed into the algorithm (see S.1.4.3). Red colors indicate high values, and hence the first major basin is a putative warm season with high water temperatures, high conductivity (σ), low river levels, and low oxygen concentrations. The second major basin (up to $4.5 \cdot 10^4$) corresponds to a putative cold season with generally inverted parameters. We can confirm these assignments by using the time annotation of the progress index shown in the bottom part of Fig. 6. These highlight that the data in the first basin indeed come from the warmest and driest months (July–September) and that the data in the second basin come from the extended winter months (November–April).

The rest of the plot reveals a few well-defined regions of homogeneous conditions that often come from specific years. These are not always well-resolved in terms of functions c or l, and one important problem contributing to this lack of resolution is lack of recurrence. This is seen most clearly for winter and spring of 2008 found at progress index values of 5 to $6 \cdot 10^4$ and indicated by linear correlation of progress index and real time. Cut values become nearly invariant, which limits the use of these annotation functions for nonrecurrent, but kinetically partitioned data. As a counterexample, the mid-summer months of 2008 found

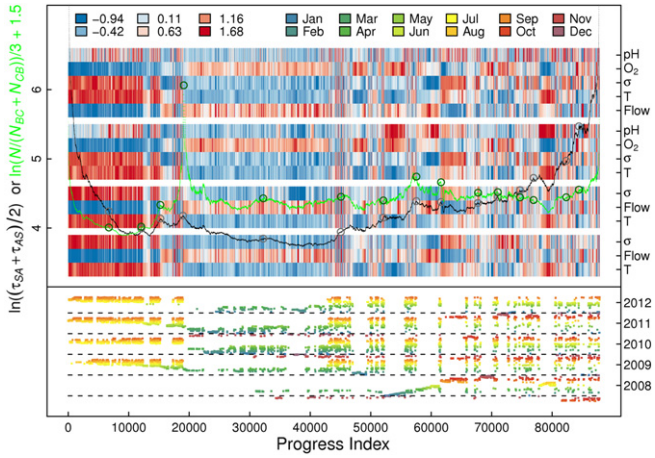


Fig. 6. Application of the exact algorithm to hydrology data. The annotation functions, c and l with a fixed $n_l = 10\,000$, derived from the progress index are plotted against the progress index as black and green curves, respectively. The data for function l were scaled and shifted as indicated in the axis label. The implied unit of time on the y-axis is a single snapshot, *i.e.*, 30 minutes. A color annotation similar to the one in Figs. 2 and 3 is shown along with these plots. Data are centered and normalized as described in the SI, S.1.4.3, and a uniform color scheme is used (legend in the upper left-hand side). Data come from four stations (that are offset visually) and encompass different measurements as indicated on the right-hand side. The lower half of the figure annotates the progress index temporally with an additional monthly color code meant to highlight the yearly patterns (legend in the upper right-hand side). Finally, circles highlight barriers identified via a measure of the locality of the progress index as described in Fig. S.4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

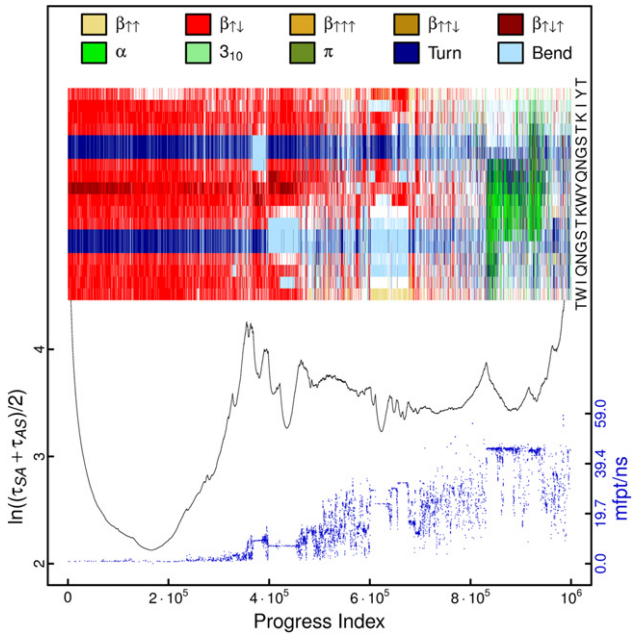


Fig. 7. Application of the approximate algorithm to Beta3S. The distance function in use is the coordinate root mean square deviation computed over backbone oxygen and nitrogen atoms of residues 3–18 after pairwise alignment. The progress index obtained with $N_g = 200$ is plotted against annotation function c. For each trajectory snapshot, we also computed DSSP annotations [35] that are presented as a color annotation (legend on top, one-letter codes for individual amino acids on the right-hand side). Only every 20th snapshot is shown to keep the size of the original vector image manageable. Lastly, we show a further kinetic annotation by plotting independent τ_{MFP} values to the native basin (small values of the progress index) for selected snapshots. The selected snapshots are the centroids of those nodes in the MSM used to generate the τ_{MFP} values, which encompass at least 10 snapshots (about 8000). Tests with values for N_g as small as 20 yielded comparable results (not shown). For plotting details, please refer additionally to the caption of Fig. S7.

at progress index values near $8 \cdot 10^4$ yield water conditions with recurrence according to both c and l. If c and l fail, it is also possible to identify barrier regions via the locality of the progress index that is known from the MST. Essentially, each snapshot is added to the set S on account of a specific edge to a specific "parent" vertex, whose position in the progress index is known. If this position is not nearby (not local), we can speculate that we have encountered a barrier region (see Fig. S.4). Putative barriers derived this way are plotted as circles in Fig. 6 and seem to offer potential in delineating basins for nonrecurrent data. Finally, a more detailed analysis is given in S.2.1. In particular, Figs. S.5 and S.6 explore differences between the exact and approximate algorithms. The latter is used exclusively for the final data set.

3.2. Reversible folding of a β-sheet miniprotein

As a final test, we apply the approximate algorithm to a complex system analyzed extensively in previous works [32,33,29]. Beta3S is a 20-residue polypeptide that undergoes reversible folding transitions at 330 K on the high ns time scale if a suitable computational model is utilized [34]. The native basin is a three-stranded β-sheet, but various other enthalpic basins are known and populated (for further details, see SI, S.1.4.2).

Fig. 7 shows representative results for the approximate progress index coupled to the simple annotation function, c. The first thing to note is the strong similarity of the plot in Fig. 7 to cut-based free energy profiles based on the same data set [32,33,29] (see also Figs. S.7 and S.8). The native basin, which the starting snapshot is part of, encompasses about 35% of the data. Secondary structure, *i.e.*, DSSP annotations [35] to the progress index are shown as well in a color plot for individual residues. These confirm the correct topology for a three-stranded β-sheet. For large values of the progress index, we find a basin comprised of an ensemble of structures rich in α-helix. In between, there is a mix of smaller, enthalpic basins that usually share part of their topology with the native state, and entropic regions without consistent order formation. Based on the DSSP annotation, it appears that function c resolves all structurally homogeneous sets of microstates suggesting that the system exhibits sufficient recurrence over the aggregate sampling time of 20 μs. This holds even for tiny basins such as the one seen at values of the progress index just past $6 \cdot 10^5$. Fig. S.8 shows the annotation with l and highlights that c provides sufficient resolution for this system.

There are two questions we want to address. First, are the resolved basins in fact kinetically homogeneous? To this extent, we constructed a network of conformational transitions based on the tree-based clustering and conformational root mean square deviations exactly as described in prior work [29] (this is also the exact same clustering used for data preorganization when generating the SST). Using a target node in the native basin as reference, we proceeded to determine the τ_{MFP} values for all other nodes. If a node contains at least 10 snapshots, the value for τ_{MFP} is plotted in Fig. 7 for all those snapshots at their respective positions in the progress index. This simple annotation confirms that – at least in reference to the native basin – the basins identified by our proposed approach are indeed kinetically homogeneous. To further address this, Fig. S.7 shows a correlation analysis of the cut-based free energy profile based on the same clustering with the results in Fig. 7. The conclusions are the same. As a corollary, a lack of kinetic homogeneity seen for example around values of the progress index of $5 \cdot 10^5$ or $8 \cdot 10^5$ correlates with parts of the profile, for which c does not indicate the presence of a basin.

The second question is with regard to the ordering of the basins by the progress index. The annotation with τ_{MFP} makes the point that there is weak correlation between a distance in the progress index and a distance in kinetics (see also Fig. S.7). This

is expected, since the sampling density in transition regions no longer represents a ruler for kinetic distance to a specific basin once multiple basins have been incorporated into set S. In analogy to cut-based free energy profiles, this also means that neighbor relations are not necessarily meaningful for larger values of the progress index as discussed in the context of Fig. 2. In summary, for this more appropriate data set compared to 3.1, the proposed scheme provides exactly the information we expected to obtain with no obvious limitations or errors in annotation.

4. Discussion and conclusions

In this contribution, we have presented a new algorithm for sorting and annotating sets of data that are the result of continuous evolution. The sorting component, *i.e.*, the progress index, is derived in both an exact form with modest computational complexity and in an approximate form that is computationally efficient and scalable to very large data sets (see Fig. 5). Such scalable algorithms are increasingly sought after due to the routine generation and storage of massive trajectories given present day computing resources [6,36,2]. The second component, *i.e.*, the various annotation functions used throughout, generally scale as $\mathcal{O}(N)$ and are of lesser cost than the progress index generation. The two components combine to yield one-dimensional plots that are able to distinguish kinetically grouped sets of microstates in complex systems that exhibit sufficient recurrence (mixing) both within and amongst basins. There are no parameters controlling size, number, or other properties of basins, and the algorithm is agnostic beyond the fact that we have to define a pairwise measure of similarity. We believe that the combination of minimal user input and high computational efficiency makes our proposed scheme a useful one.

The total runtime for generating Fig. 7 was on the order of minutes for a trajectory of 10^6 snapshots. This highlights the utility of the approach in quickly and reliably partitioning a complex system into an annotated set of basins. We are unaware of alternative methods offering comparable amounts of information at this cost. The strengths of the approach rest on the use of all snapshots, *i.e.*, the lack of any binning or other *a priori* grouping (the auxiliary clustering is for efficiency only (see 2.3), and has no direct bearing on the results (see Fig. S.6)). The kinetic annotation functions, c and l (see 2.1 and 2.2), operate relative to the time resolution of the data and will correctly lump all snapshots together if the latter is too coarse (see Figs. 3 and S.2). Actual failure is possible if small basins are placed in regions of high inherent curvature (see Fig. 3). This is an issue of the signal-to-noise ratio, and we expect it to be corrected by increasing the amount of data or using a different starting point. Any lack of recurrence is a potentially more critical issue and is encountered in Fig. 6. However, it need not result from non-stochastic evolution of the system, but can also result from an inappropriately high dimensionality in representation. In the latter case, the point density becomes so low everywhere that basins are no longer identifiable.

The last comment above implies that the utility of data processing algorithms of this type rests on the appropriateness of the distance function. This is a very fundamental problem, but there is little rigorous work comparing combinations of different classes of distance functions coupled to different representations of a complex system [37]. A more active and closely related area of research is that of finding suitable reaction coordinates for complex systems that preserve correct, coarse-grained kinetics and thermodynamics [38,33,39,40]. Viewed as a simple grouping scheme [23], our approach offers the advantage over the majority of algorithms that there is no parameter controlling the number or size of clusters. Moreover, comparable groupings are normally the result of a two-stage process: efficient, fine-grained clustering is followed by suitable refinement [41]. Our approach shares a strong formal similarity with the OPTICS clustering algorithm that also utilizes a combination of sorting and annotation [42]. We emphasize that few algorithms in this class operate at such low time complexity, *e.g.*, [43,44,29]. The reliance on geometric continuity during system evolution is shared explicitly with methods computing eigenvectors of a kernel-based density estimate given the full $\mathcal{O}(N^2)$ Laplacian matrix, *i.e.*, diffusion maps [15,39]. These methods not only require choosing a kernel function (or at least parameter(s) for it), but the reliance on the Laplacian matrix renders them infeasible for data sets exceeding $\sim 10^5$ snapshots. Lastly, we briefly mention path sampling approaches. With suitably chosen end points, these methods can yield comparable information [45–48], because they directly probe kinetic connectivity of different sets of microstates. Of course, they are conjoined with the sampling protocol itself, *i.e.*, they are not pure analysis schemes applicable to any continuous data set, and require significant human input. This is also true for metadynamics [49] and many related approaches, *e.g.*, a recent approach to sequential basin discovery [50].

The algorithm as described here has been implemented in the CAMPARI software package [51], and the current development version is available from the authors on request (campari.software@gmail.com). Ongoing work is targeting three areas. First, can we automatize feature selection using an appropriate criterion of optimality, *i.e.*, is it possible to eliminate the need to manually define a distance function? Second, for the localized cut function, l, is there an iterative, but efficient procedure that determines a suitable value of $n_l(i)$ for all snapshots? The current restriction to one or a few values of n_l clearly lacks general utility. Third, can we identify additional annotation functions that can be quantitatively related to relevant time scales of the system? We believe that addressing these questions opens up fruitful avenues for future research toward routine analysis of large data sets continuous in time.

Acknowledgments

We thank Dr. Ting Zhou for sharing data on Beta3S used for runtime analysis [52]. This work was supported by a grant of the Swiss National Science Foundation to A. C.

Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.cpc.2013.06.009.

References

[1] I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic, P. Canal, D. Casadei, O. Couet, V. Fine, L. Franco, G. Ganis, A. Gheata, D. Gonzalez Maline, M. Goto, J. Iwaszkiewicz, A. Kreshuk, D. Marcos Segura, R. Maunder, L. Moneta, A. Naumann, E. Offermann, V. Onuchin, S. Panacek, F. Rademakers, P. Russo, M. Tadel, ROOT—A C++ framework for petabyte data storage, statistical analysis and visualization, Comput. Phys. Comm. 180 (2009) 2499–2512.
[2] D. Hasenkamp, A. Sim, M. Wehner, K. Wu, Finding tropical cyclones on a cloud computing cluster: using parallel virtualization for large-scale climate simulation analysis, in: J. Qiu, G. Zhao, C. Rong (Eds.), 2010 IEEE Second International Conference on Cloud Computing Technology and Science, Indianapolis, IN, USA, November 30–December 3, 2010, CloudCom, IEEE Computer Society Conference Publishing Services, Los Alamitos, CA, USA, 2010, pp. 201–208.
[3] E.E. Schadt, M.D. Linderman, J. Sorenson, L. Lee, G.P. Nolan, Computational solutions to large-scale data management and analysis, Nature Rev. Genet. 11 (2010) 647–657.
[4] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science 334 (2011) 517–520.
[5] G. Settanni, F. Rao, A. Caflisch, Φ-value analysis by molecular dynamics simulations of reversible folding, Proc. Natl. Acad. Sci. USA 102 (2005) 628–633.
[6] V. Springel, S.D.M. White, A. Jenkins, C.S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J.A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg, F. Pearce, Simulations of the formation, evolution and clustering of galaxies and quasars, Nature 435 (2005) 629–636.

[7] Y. Wang, J.Y.-J. Shyy, S. Chien, Fluorescence proteins, live-cell imaging, and mechanobiology: seeing is believing, Annu. Rev. Biomed. Eng. 10 (2008) 1–38.
[8] B.P. Kirtman, C. Bitz, F. Bryan, W. Collins, J. Dennis, N. Hearn, J.L. Kinter III, R. Loft, C. Rousset, L. Siqueira, C. Stan, R. Tomas, M. Vertenstein, Impact of ocean model resolution on CCSM climate simulations, Clim. Dyn. 39 (2012) 1303–1328.
[9] N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. 438 (2007) 237–329.
[10] I.G. Kevrekidis, C.W. Gear, G. Hummer, Equation-free: the computer-aided analysis of complex multiscale systems, AIChE J. 50 (2004) 1346–1355.
[11] S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, U. Alon, Coarse-graining and self-dissimilarity of complex networks, Phys. Rev. E 71 (2005) 016127.
[12] J.D. Halley, D.A. Winkler, Classification of emergence and its relation to self-organization, Complexity 13 (2008) 10–15.
[13] M. Sips, B. Neubert, J.P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency, Comput. Graph. Forum 28 (2009) 831–838.
[14] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, Pattern Recognit. 41 (2008) 176–190.
[15] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmon. Anal. 21 (2006) 113–127.
[16] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[17] S.V. Krivov, M. Karplus, One-dimensional free-energy profiles of complex systems: progress variables that perserve the barriers, J. Phys. Chem. B 110 (2006) 12689–12698.
[18] F. Noé, I. Horenko, C. Schütte, J.C. Smith, Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states, J. Chem. Phys. 126 (2007) 155102.
[19] J.D. Chodera, N. Singhal, V.S. Pande, K.A. Dill, W.C. Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics, J. Chem. Phys. 126 (2007) 155101.
[20] S. Muff, A. Caflisch, Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein, Proteins: Struct. Funct. Bioinform. 70 (2008) 1185–1195.
[21] J.M. Carr, D.J. Wales, Folding pathways and rates for the three-stranded β-sheet peptide Beta3s using discrete path sampling, J. Phys. Chem. B 112 (2008) 8760–8769.
[22] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (2009) 056117.
[23] R. Xu, D. Wunsch II, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (2005) 645–678.
[24] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. 19 (1997) 153–158.
[25] B. Keller, X. Daura, W.F. van Gunsteren, Comparing geometric and kinetic cluster algorithms for molecular simulation data, J. Chem. Phys. 132 (2010) 074110.
[26] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, F. Cordes, From simulation data to conformational ensembles: structure and dynamics-based methods, J. Comput. Chem. 20 (1999) 1760–1774.
[27] D.J. Wales, Energy landscapes: some new horizons, Curr. Opin. Struct. Biol. 20 (2010) 3–10.
[28] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Modern Phys. 81 (2009) 591–646.
[29] A. Vitalis, A. Caflisch, Efficient construction of mesostate networks from molecular dynamics trajectories, J. Chem. Theory Comput. 8 (2012) 1108–1120.
[30] M. Daszykowski, B. Walczak, D.L. Massart, Projection methods in chemistry, Chemometr. Intell. Lab. Syst. 65 (2003) 97–112.
[31] J. Nešetřil, E. Milková, H. Nešetřilová, Otakar Borůvka on minimum spanning tree problem, translation of both the 1926 papers, comments, history, Discrete Math. 233 (2001) 3–36.
[32] S.V. Krivov, S. Muff, A. Caflisch, M. Karplus, One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: new insights into the folding process, J. Phys. Chem. B 112 (2008) 8701–8714.
[33] B. Qi, S. Muff, A. Caflisch, A.R. Dinner, Extracting physically intuitive reaction coordinates from transition networks of a β-sheet miniprotein, J. Phys. Chem. B 114 (2010) 6979–6989.
[34] P. Ferrara, J. Apostolakis, A. Caflisch, Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations, J. Phys. Chem. B 104 (2002) 5000–5010.
[35] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[36] D.E. Shaw, M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, J.C. Chao, M.P. Eastwood, J. Gagliardo, J.P. Grossman, C.R. Ho, D.J. Ierardi, I. Kolossváry, J.L. Klepeis, T. Layman, C. McLeavey, M.A. Moraes, R. Mueller, E.C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S.C. Wang, Anton, a special-purpose machine for molecular dynamics simulation, in: D. Tullsen, B. Calder (Eds.), Proceedings of the 34th Annual International Symposium on Computer Architecture, San Diego, CA, USA, June 9–13, 2007, ISCA'07, ACM, New York, NY, USA, 2007, pp. 1–12.
[37] P. Cossio, A. Laio, F. Pietrucci, Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? Phys. Chem. Chem. Phys. 13 (2011) 10421–10425.
[38] R.B. Best, G. Hummer, Reaction coordinates and rates from transition paths, Proc. Natl. Acad. Sci. USA 102 (2005) 6732–6737.
[39] M.A. Rohrdanz, W. Zheng, M. Maggioni, C. Clementi, Determination of reaction coordinates via locally scaled diffusion map, J. Chem. Phys. 134 (2011) 124116.
[40] S.V. Krivov, Is protein folding sub-diffusive? PLoS Comput. Biol. 6 (2010) e1000921.
[41] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, in: J. Widom (Ed.), SIGMOD'96: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, QC, Canada, June 4–6, 1996, ACM Press, New York, NY, USA, 1996, pp. 103–114.
[42] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, in: J. Clifford, R. King (Eds.), SIGMOD'99: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, PA, USA, May 31–June 03, 1999, ACM Press, New York, NY, USA, 1999, pp. 49–60.
[43] A. Hinneburg, D.A. Keim, Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering, in: M.P. Atkinson, M.E. Orlowska, P. Valduriez, S.B. Zdonik, M.L. Brodie (Eds.), Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, September 7–10, 1999, Morgan Kaufmann, San Francisco, CA, USA, 1999, pp. 506–517.
[44] R.L.F. Cordeiro, A.J.M. Traina, C. Faloutsos, C. Traina Jr., Halite: fast and scalable multi-resolution local-correlation clustering, IEEE Trans. Knowl. Data Eng. 25 (2013) 387–401.
[45] P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler, Transition path sampling: throwing ropes over rough mountain passes, in the dark, Annu. Rev. Phys. Chem. 53 (2002) 291–318.
[46] D.J. Wales, Discrete path sampling, Mol. Phys. 100 (2002) 3285–3305.
[47] W. E, W. Ren, E. Vanden-Eijnden, Simplified and improved string method for computing the minimum energy paths in barrier-crossing events, J. Chem. Phys. 126 (2007) 164103.
[48] P. Faccioli, Characterization of protein folding by dominant reaction pathways, J. Phys. Chem. B 112 (2008) 13756–13764.
[49] A. Laio, M. Parrinello, Escaping free-energy minima, Proc. Natl. Acad. Sci. USA 99 (2002) 12562–12566.
[50] Y.V. Sereda, A.B. Singharoy, M.F. Jarrold, P.J. Ortoleva, Discovering free energy basins for macromolecular systems via guided multiscale simulation, J. Phys. Chem. B 116 (2012) 8534–8544.
[51] A. Vitalis, A. Steffen, N. Lyle, A.H. Mao, R.V. Pappu, Campari v1.0, http://sourceforge.net/projects/campari (accessed 30.10.12).
[52] T. Zhou, A. Caflisch, Free energy guided sampling, J. Chem. Theory Comput. 8 (2012) 2134–2140.

Supplementary Information to
A Scalable Algorithm to Order and Annotate Continuous Observations Reveals the Metastable States Visited by Dynamical Systems

Nicolas Blöchliger[a], Andreas Vitalis[a,*], Amedeo Caflisch[a]

[a]*Department of Biochemistry*
University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland

Abstract

Advances in IT infrastructure have enabled the generation and storage of very large data sets describing complex systems continuously in time. These can derive from both simulations and measurements. Analysis of such data requires the availability of scalable algorithms. In this contribution, we propose a scalable algorithm that partitions instantaneous observations (snapshots) of a complex system into kinetically distinct sets (termed basins). To do so, we use a combination of ordering snapshots employing the method's only essential parameter, *i.e.*, a definition of pairwise distance, and annotating the resultant sequence, the so-called progress index, in different ways. Specifically, we propose a combination of cut-based and structural annotations with the former responsible for the kinetic grouping and the latter for diagnostics and interpretation. The method is applied to an illustrative test case, and the scaling of an approximate version is demonstrated to be $O(N \log N)$ with N being the number of snapshots. Two real-world data sets from river hydrology measurements and protein folding simulations are then used to highlight the utility of the method in finding basins for complex systems. Both limitations and benefits of the approach are discussed along with routes for future research.

Keywords: Complex System, Trajectory Analysis, Scalable Algorithm, Minimum Spanning Tree, Free Energy Basins

*To Whom Correspondence Should be Addressed
Email addresses: n.bloechliger@bioc.uzh.ch (Nicolas Blöchliger), a.vitalis@bioc.uzh.ch (Andreas Vitalis), caflisch@bioc.uzh.ch (Amedeo Caflisch)

S.1. Supplementary Methods

S.1.1. Implementation of Exact Algorithm

The implementation described in the following is used to provide an algorithm for the following scheme as proposed in the main text:

Choose a starting snapshot $s_1 \in T$ and create the set $S_1 = \{s_1\}$. Initialize the cut function, $c : \{1, \ldots, N\} \to \mathbb{N}$, to 2. Then, for $i = 1, \ldots, N-1$ do the following:

1. Define s_{i+1} as the snapshot in $T \setminus S_i$ realizing the minimum of $d(\cdot, S_i) = \min_{j=1,\ldots,i} d(\cdot, s_j)$.
2. Let $S_{i+1} = S_i \cup \{s_{i+1}\}$.
3. Define $c(i+1) = \sum_{j=1}^{N-1} \zeta_{S_{i+1}}(t_j, t_{j+1})$.

Here, function ζ is given by eq. 1 in the main text. The exact **progress index** of T starting with s_1 is defined as the sequence $S(T, s_1) = (s_1, \ldots, s_N)$. Each entry i is associated with a value for the cut function, $c(i)$.

In order to guarantee the scheme to be exact, we theoretically need to know all $O(N^2)$ pairwise distances. Then, we can - for each snapshot - create a list containing all other snapshots ordered by distance along with the distance. This already has a complexity of $O(N^2 \log N)$ on account of the required sorting of a list of size $O(N)$ for every snapshot. From this set of ordered lists, the progress index is conveniently generated in $O(N^2)$ time. This is because for iteration i in the above scheme, we need to scan only the nearest, eligible neighbor for each current member of the set S, *i.e.*, we find the minimum one of $i = O(N)$ candidates. To keep track of eligibility, we utilize a pointer array to the smallest, eligible entry in each list, and updating the pointer array again has limiting complexity of $O(N)$ for each iteration.

Suppose now we assume that the density of points is homogeneous enough such that the longest edge in the underlying minimum spanning tree (MST) has a value of d_{max} that is substantially smaller than the majority of distance values found in the ordered lists. Then, the following heuristic emerges. With a reasonable guess of d_{max}, we first use an efficient clustering algorithm with controllable cluster size to find a set of cluster centroids. For each cluster k, we can compute the maximum distance from its centroid as r_{c_k}. Then, by virtue of d being a metric, we can exclude all pairwise distance comparisons for snapshots belonging to clusters k and m, whose centroids are further apart than $d_{max} + r_{c_k} + r_{c_m}$. This reduces the computational cost of the implementation twofold: first, not all $O(N^2)$ distances are evaluated (in our tests on n-butane, the fraction evaluated ranged from 5-30%); second, the sorting of the lists truncated to d_{max} needs less than $O(N^2 \log N)$ time for each list. It is important to point out that we usually expect the required value for d_{max} to decrease with increasing sampling density meaning that the overall complexity may be reduced to $O(N^2)$ or less.

Unfortunately, there are three problems associated with the above scheme. First, the required value of d_{max} is difficult to guess. To obtain a MST, the algorithm may have to be rerun a few times with increasing values for d_{max}. Second, the heuristic in use is dependent on the structure of the data, *i.e.*, it is not universally applicable. It is straightforward to construct pathological cases, in which a single edge of the MST is so much longer than the rest that the truncated lists are nearly as long and as expensive to compute as the complete $O(N)$ lists. Third, even if the data conform to the assumptions of the heuristic, the memory required for storing the lists still grows superlinearly with N. This is in contrast to the number of edges of the MST that is $N-1$.

With the lists generated, the efficiency of the generation of the progress index itself can potentially be improved by first combining all distances and snapshot pairs in the truncated lists to a single list that is then globally sorted by distance. Obviously, the complexity of this operation is favorable compared to the implementation described above only if the number of items is significantly smaller than $(N-1)N/2$. From the globally sorted list, we can derive the MST via Kruskal's algorithm [1] with lower time complexity due to the edges already being sorted. For data sets for n-butane, we found an effective scaling exponent of computational cost over a range of 900–90000 snapshots of 1.75.

S.1.2. Derivation of Combinatorial Prior for Cut Function

As shown in Fig. 3 of the main text, a progressive decrease in temporal resolution eventually yields a default profile with parabolic shape. If the saving frequency exceeds all relevant time scales of the system, the trajectory essentially looks random, and the annotation function c reports what looks like a single basin. However, there are combinatorial reasons for why the result is not flat along the progress index, and these reasons are treated explicitly next.

The idea of the analytical derivation centers around the number of different ways a trajectory can be randomly partitioned into two sets, S_i and $A_i = T \setminus S_i$. For given $i = |S_i|$ and $j = |A_i| = N - i$, there are

$$r_c = 2\binom{i-1}{q}\binom{j-1}{q}$$

trajectories with a value for c of $1 \leq c = 2q + 1 \leq 2\min\{i, j\} - 1$, and there are

$$r_c = \binom{i-1}{q}\binom{j-1}{q-1} + \binom{j-1}{q}\binom{i-1}{q-1}$$

trajectories with a value for c of $2 \leq c = 2q \leq 2\min\{i, j\}$. For a trajectory whose snapshots are randomly assigned to S_i and A_i (satisfying $|S_i| = i$), the probability that its cut value is c is given by

$$p_c = r_c \Big/ \binom{N}{i}.$$

Using Vandermonde's identity, we get the expected c_{prior} as the following expectation value:

$$\mathbb{E}(c(i)) = \sum_{c=1}^{2\min\{i,j\}} c p_c = 2i \binom{N-1}{i} \Big/ \binom{N}{i} = 2i(N-i)/N.$$

The function defined above is used in Fig. 3 of the main text (dashed line), and corresponds to a combinatorial prior function that is a direct result of the differing asymmetry in partition sizes for different values of the progress index. The resulting profile is symmetric around the point $N/2$ and is independent of the length of the trajectory in the sense that $\mathbb{E}(c_{\lambda N}(\lambda i))/\lambda N = \mathbb{E}(c_N(i))/N$ for all $\lambda \in \mathbb{N}$ (here, the subscript denotes the length of the trajectory).

S.1.3. Implementation Details for the Approximate Algorithm

Recalling the exact algorithm conceptually (see Section S.1.1), we can divide the task into two components, *i.e.*, construction of the MST and parsing of the MST to yield the annotated progress index. For the latter, for typical data sets, the computational cost will scale linearly with N given a spanning tree. In contrast, the former is prohibitively expensive for large data sets as described above. Assuming we know the ordered list of nearby snapshots for every snapshot, Borůvka's well-known algorithm [2] assembles the MST by successively joining subtrees in $O(N \log N)$ time. Therefore, approximations are introduced that are meant to replace the use of an ordered list of all other snapshots with a set of unordered and nearby snapshots and of controlled size, which allows any given merging operation to happen in constant or nearly constant time. The result is no longer an MST, but rather a short spanning tree (SST).

Specifically, this happens by data preorganization and random guessing with a fixed maximum number of guesses, which corresponds to the parameter N_g used throughout. Data preorganization allows defining a list of snapshots, τ_i, that contains candidates with small values of the distance to the set S_i describing a subtree at any given stage of Borůvka's algorithm. The notion of "small" is understood qualitatively in relation to the distribution of this quantity for all snapshots in $T \setminus S_i$. τ_i can be assembled by clustering the data set prior to the construction of the spanning tree. Specifically, τ_i is the list of unique snapshots constructed from all clusters that S_i spans into. Following the algorithm, it becomes clear that eventually all cluster members will be exhausted preventing further merging steps of subtrees. This is where the idea of a hierarchical clustering becomes critical. Hierarchical data preorganization implies that we obtain a clustering for a series of chosen resolutions of increasing coarseness. If clusters are exhausted at the finest resolution, τ_i is simply assembled at the next coarser level that yields a nonempty set.

In principle, any hierarchical clustering algorithm that does not generate cluster overlap and reflects local density could be used provided that it operates in at most $O(N \log N)$ time with data set size and in linear time with data dimensionality. In our implementation, we use a recently developed top-down, tree-based clustering algorithm meeting these requirements [3]. The resultant hierarchical tree of clusters is not to be confused with the MST or SST at the snapshot level considered here. While the reader is referred to the literature for details, a brief summary is as follows. The clustering algorithm relies on one main (a minimum threshold distance, t_1) and two auxiliary parameters (the tree height, H, and a maximum threshold distance, t_H). From the top to the bottom level the data set is clustered with increasing resolution such that parent-child relationships defining the tree of clusters can be exploited to achieve near linear scaling with data set size. Each of the H tree levels is associated with a threshold distance t_k determined by linear interpolation between t_H and t_1. The data is processed sequentially. Starting on the top level, snapshot j is added to its nearest cluster on level H if the distance does not exceed t_H, otherwise j spans a new cluster on its own. Then for each level k down to level 2, j is added to the nearest cluster on this level provided that the distance does not exceed t_k, otherwise it spans a new cluster. The key steps guaranteeing efficiency are that 1) only the children of the cluster of j on level $k+1$ are scanned, and that 2) distances of snapshots to clusters are measured as distances to the centroid of the cluster. Using simple algebra, these centroids can be updated continuously without additional cost pending that the distance function in use is Euclidean. The first pass creates a "raw" tree with minor errors caused primarily by centroid drift. Therefore, in a second pass of the data, centroids at all populated levels are kept fixed, and snapshots are simply reassigned to clusters. In addition, the clusters at the leaf level (finest resolution) are now created, which implies that the leaf level results are of higher quality. This is also a reasonable property for SST construction given that the majority of SST edges are expected to derive from neighbor relations encoded in leaf level clusters.

With the restricted list of snapshots, τ_i, in hand, a second approximation may be introduced. Specifically, if the size of τ_i exceeds the value of the parameter N_g, N_g candidates are picked randomly, and the one yielding the shortest edge becomes a putative edge of the SST; otherwise, the search is exhaustive over the set of candidate edges, and the approximation is purely at the level of reducing the search space to τ_i. Clearly, the second approximation may be severe. Consider a case where τ_i is constituted from multiple, coarse clusters spanning a large volume in data space. The distribution of points in τ_i in relation to the members of S_i may be

heavily skewed, and/or the ratio $N_g/|\tau_i|$ may be unfavorable. In either case, the likelihood of introducing a significantly inaccurate neighbor relationship is large. These types of issues imply that it is not straightforward to optimize the hierarchical clustering for SST construction, and we have not attempted to do so rigorously. The most important property is probably that the leaf clusters be tight, free of overlap, and somewhat matched in size to the choice for N_g.

In addition, there are some technical points to consider. First, the assembly of snapshot lists for each subtree must be handled efficiently by using dedicated, but straightforward data structures such as different types of linked lists. Memory is allocated dynamically, but with sufficient buffering to prevent slowdown by frequent allocation events. A heuristic is used to determine whether to use partial snapshot lists created during clustering or whether to recreate the list for a given cluster. Second, larger subtrees will eventually span multiple clusters. It may happen that one or more, but not all of these clusters are exhausted at a given level requiring a jump to the next coarser resolution. In this case, τ_i remains restricted to the finest level for which any cluster still contains eligible snapshots. This strategy is meant to exploit the fact that neighbor relationships will be most meaningful at the finest resolution (leaf) level. Third, note that at each step of the algorithm there is only a single minimum distance considered for every subtree, and that the corresponding edge is added to the SST only if it does not introduce a cycle. Cycles are avoided by updating an index array denoting subtree memberships for all snapshots immediately after each merging event.

S.1.4. Data Sets

S.1.4.1. n-butane

We used torsional space, stochastic dynamics simulations in the gas phase at 400 K to obtain trajectories of varying length. In all cases, only a single molecule of *n*-butane was present, the integration time step was 5 fs, and the number of trajectory snapshots that were analyzed was 30000. The only term to the potential energy were the torsional potentials native to the OPLS-AA force field [4]. The chosen representation consisted of the three dihedral angles themselves with appropriate corrections for computing Euclidean distances between snapshots [3].

S.1.4.2. Beta3S Miniprotein

Data were taken from equilibrium sampling of the polypeptide Thr-Trp-Ile-Gln-Asn-Gly-Ser-Thr-Lys-Trp-Tyr-Gln-Asn-Gly-Ser-Thr-Lys-Ile-Tyr-Thr with the terminal residues in zwitterionic state. At least in the limits of a specific continuum description of solvation [5], this peptide undergoes reversible folding transitions [6, 7] between a well-ordered, three-stranded β-sheet conformation and a coil-like unfolded state ensemble. This is augmented by various enthalpically stabilized, non-native basins, most prominently a partially ordered ensemble of α-helix rich conformations.

The simulation data were obtained from prior work [8], and a data set of size $8 \cdot 10^6$ trajectory snapshots was considered for analysis. The pairwise, Euclidean distance function is defined on a chosen representation that here consisted of 273 interatomic distances between backbone nitrogen and oxygen atoms. This representation deemphasizes fast degrees of freedom such as side chain rotamer states. For the purpose of runtime analysis (Fig. 5A) of the approximate algorithm, the entire $8 \cdot 10^6$ snapshots were read with increasing skip to arrive at data sets of reduced size. In order to make the results comparable, we kept the number of guesses, N_g, the tree height, and the maximum threshold criterion constant at values of 20, 16, and 8 Å, respectively. To obtain a constant average cluster size, the finest threshold value was set to 1.82, 1.67, 1.52, 1.35, 1.15, 1.03, 0.88, and 0.76, for increasing snapshot numbers from 62500 to $8 \cdot 10^6$. It should be noted that this parameter is also the only one that shows a weakly systematic impact on the total weight of the SST, which is a measure of the quality of the approximation (see Fig. S.5). For Fig. 5C we used a data set of fixed size corresponding to the case with $N = 10^6$ in Fig. 5A. N_g was varied to investigate runtime dependency on this parameter. Quantifying the influence of the dimensionality of representation is more complicated, and we chose to first transform the 273 interatomic distances into principal components sorted by decreasing total variance (see Fig. 5B). For the full dimensionality, this has no impact on the results of the clustering or the SST construction. It does, however, allow a more straightforward reduction in dimensionality by simply discarding more and more of those dimensions with the smallest variance, which are presumed to encapsulate the least information.

For the data in Fig. 7, we used data on the identical system and physical model, but from a different set of simulations [7]. This is meant to facilitate comparisons to published work [7, 9, 3]. The data set is comprised of $N = 10^6$ snapshots saved at an interval of 20 ps, and the representation consists of the Cartesian coordinates of the backbone nitrogen and oxygen atoms of residues 3–18. A pairwise distance is defined as the root mean square deviation of atomic coordinates after pairwise alignment. It should be pointed out that the inclusion of translation and rotation operators poses technical challenges in the hierarchical clustering underlying the approximate approach as discussed [3]. The clustering used a tree height of 16, and a maximum and minimum threshold radius of 10.0 and 1.5 Å, respectively (identical to Fig. 6A in [3]). It yielded 161778 clusters, and the resultant network of conformational transitions was used to derive the τ_{MFP} annotations in Figs. 7, S.7, and S.8 as well as the cut-based free energy profile in Fig. S.7.

S.1.4.3. Hydrology Data Set

Because of storage constraints, it is difficult to find non-synthetic data on an accessible topic hosted on public servers such that there are continuous recordings of quantities or parameters with both time resolution and recurrence that allows one to make

statements regarding basins. Here, we have chosen river hydrology parameters (temperature in °C, pH (unfiltered), specific conductance in µS/cm at 25°C, discharge in cubic feet per second, and dissolved oxygen in mg/ml) available from the following stations in Oregon, USA:

Site number	River	Coordinates	Altitude
USGS 14211010	Clackamas River	45°22'46"N 122°34'34"W	0 ft.
USGS 14209710	Clackamas River	45°10'02"N 122°09'18"W	840 ft.
USGS 14138870	Fir Creek	45°28'49"N 122°01'28"W	1440 ft.
USGS 14138850	Bull Run River	45°29'54"N 122°00'40"W	1080 ft.

Table S1: The first column lists the station ID within the system of the United States Geological Survey (USGS) [10]. For each of the four stations, we here list the river (column 2) it measures along with complete geographic coordinates as latitude, longitude, and altitude (columns 3–4) [10]. The stations are listed in the same order as their data are shown in Figs. 6 and S.6. All these rivers are ultimately indirect tributaries to the Columbia river, which drains into the Pacific Ocean. The Portland area has mild, wet winters, comparatively cool summers, and is classified to be part of the cool, dry-summer subtropical (Csb) zone in the Köppen–Geiger classification system [11]. The rivers form part of the system that is relevant to the freshwater supply of the Portland metropolitan area and to the generation of hydroelectric power. In 2008, the lower basin of the Clackamas river was subject to a USGS investigation regarding river pollution from pesticides and herbicides based on data from years 2000–2005 [12].

These hydrology data are available with a temporal resolution of 30 minutes and over a period of about 5 years (from October 2007 to the present). In some cases, homogenization of the time axis required shifts of a few minutes for the actual time of measurement. Due to malfunctioning equipment, severe weather, or scheduled outages, data are incomplete (ca. 2.6% of points). In such cases, we interpolated linearly between the two measured data points bracketing a stretch of missing data. This is reasonable even for missing stretches of multiple days since river hydrology data are neither prone to strong random fluctuations nor to pronounced diurnal patterning. After homogenization and completion of the data, uniform noise was added to compensate for the lack of resolution in measurements. The width of the noise function was centered at the measured value and chosen in accordance with the dominant apparent resolution for each type of measurement. Discharge (streamflow) data were converted to logarithmic space before centering all the data. To achieve similar impact of each dimension, data were then normalized by their apparent standard deviations. These complete, centered, and normalized data are what is shown as color annotations in Fig. 6 of the main text and Fig. S.6.

S.2. Supplementary Results

S.2.1. Hydrology Data for Rivers Near Portland, Oregon

As outlined in the main text, we use the hydrology data (see S.1.4.3) for two main purposes: 1) highlighting data-dependent difficulties in applying the algorithm; 2) demonstrating the algorithm's utility on a real-world data set. There are some finer details regarding both points that are presented here instead of in 3.1 in the main text. To preserve clarity, some results are repeated here.

Figs. 6 and S.6 both reveal two major basins corresponding to warm and cold seasons, respectively. The cold season is the more heterogeneous of the two, and this is manifested predominantly by a broader range of discharge (flow) levels. 2008 appears to have been a year with anomalous conditions and is largely excluded from both major basins. The rest of the plot is partially comprised of a number of "entropic" regions constituted by mixed conditions from throughout the year. This is probably an aspect specific to data following an annual rhythm that generally cycles between two sets of extreme values, and it is expected that fall and spring are overrepresented in these "entropic" regions. The remainder are well-defined regions of homogeneous conditions that often come from specific years. As outlined in the main text, these tend to be resolved rather poorly in terms of the cut functions c or l on account of a lack of recurrence. Using the example of the winter and spring of 2008 found at progress index values of 5 to $6 \cdot 10^4$ in Fig. 6, we note that the conditions are unique with a very high pH at site #3 and very low water temperatures in winter. The linear correlation of progress index and real time indicates poor recurrence. This is because adding snapshots in their exact temporal sequence will leave the cut function invariant, *i.e.*, the number of transitions between sets S_i and $T \setminus S_i$ is constant for a range of consecutive i. Difficulties notwithstanding, the method allows identification of 2008 as a year with an unusually cold first half and friendly and dry weather deep into fall (see values for the progress index around $6.3 \cdot 10^4$) [13].

As alluded to in the main text, we also explore an alternative approach to the identification of barrier regions. This approach utilizes the locality of the progress index, *i.e.*, the difference in progress index position between a snapshot $i + 1$ and the snapshot in set S_i that it shares an MST edge with. This is discussed in detail in Fig. S.4 and its caption. With a suitable amount of averaging, this produces a plot that highlights putative barrier regions. These are then plotted as circles in Fig. 6. The aforementioned winter and spring basins of 2008 at progress index values of 5 to $6 \cdot 10^4$ are a good example for the utility of this approach. Specifically, the cut functions do not allow delineation of the winter basin from the data immediately to the left, whereas an identification via

nonlocality of the progress index is successful. In summary, these results emphasize the need to combine several annotations to extract meaningful information from a challenging data set.

As a final test, we want to utilize this realistic data set to evaluate how much the results depend on the spanning tree used to generate the progress index. Specifically, we are interested in identifying potential problems with the approximate algorithm that uses a parameter-dependent SST in place of the MST. First, we consider the SST on its own. In this context, the total weight of the spanning tree is a useful quantifier, given that this quantity is minimal for the MST. Fig. S.5 shows a comparative analysis for various choices of the number of guesses, N_g, used to construct the SST (see 2.3 in the main text and S.1.3 above). Clearly, N_g can be used to systematically decrease the weight of the SST. However, the value of the MST in not reached in an asymptotic manner, which must be on account of search restrictions introduced by data preorganization (see S.1.3). This means that the influence of the implied approximation on both progress index and annotation functions is difficult to predict. Fig. S.6 and its caption describe an example application of the approximate algorithm to the hydrology data. The conclusion is that, for this particular example, the SST actually provides better resolution in terms of the annotation functions, because it introduces artificial recurrence within basins. This comes at a cost, however, that is manifested in increased variability between plots starting from different snapshots and between plots starting from the same snapshot for different SSTs (not shown).

S.3. Supplementary Figures

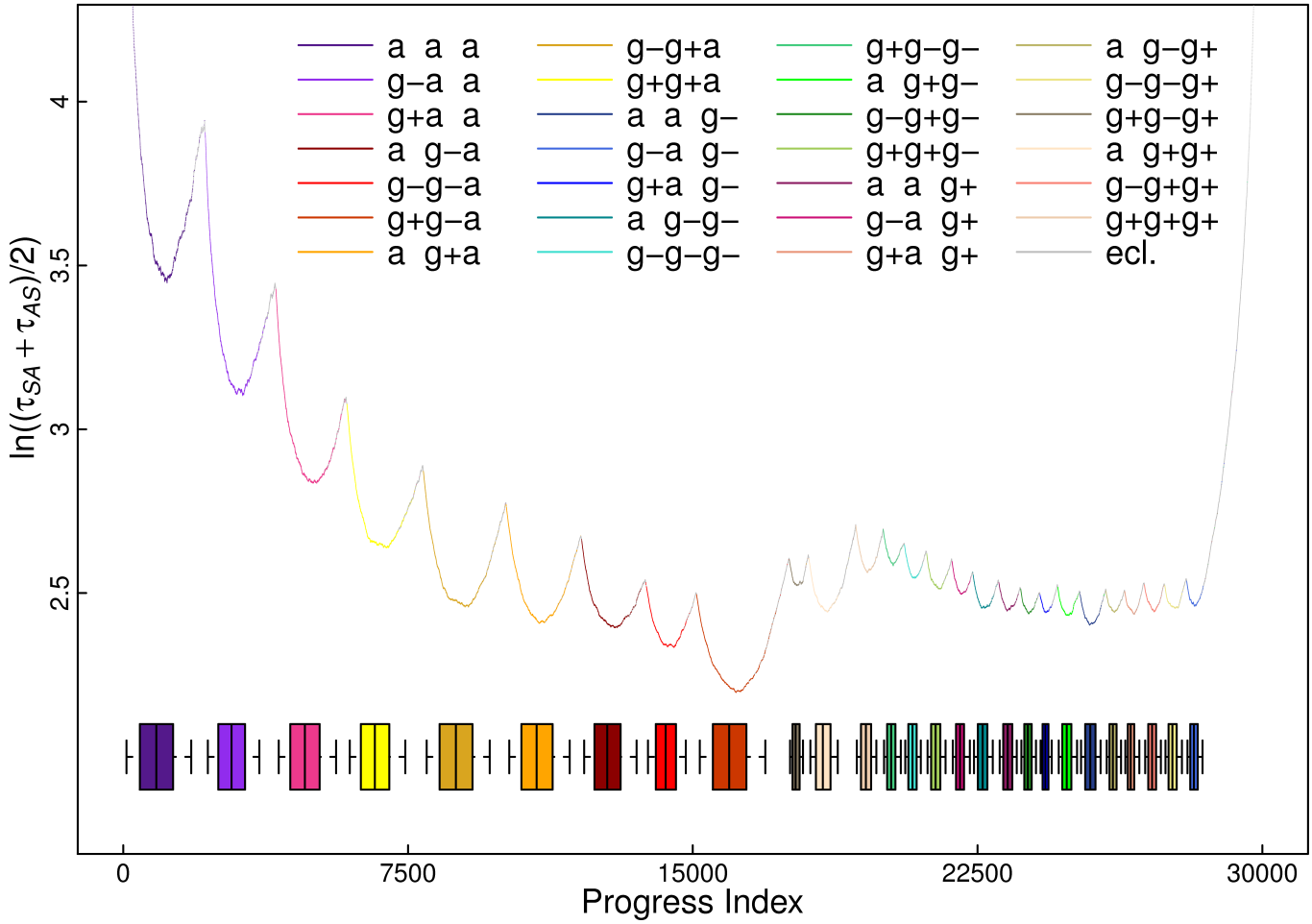


Figure S.1: Illustration of the approach using *n*-butane. This figure is largely similar to Fig. 2 in the main text. Our proposed approach yields a curve that resolves all 27 basins of the system. Microstates are annotated by color, and box plots are shown that quantify the distribution of snapshots annotated via the dihedral-angle based binning. For each basin, a box is drawn indicating the interval that the central 50% of the snapshots belonging to that basin are confined to. Whiskers indicate the central 90% of the snapshots in that basin. Medians are shown as black, vertical lines. Compared to the maxima in the curve, boxes and medians appear skewed to the left. This emphasizes that within each basin eclipsed microstates are concentrated toward the right (larger progress index), which is a natural result of the way the progress index is constructed. This is also qualitatively apparent from the gray dots indicating such eclipsed microstates. The implied unit of time on the y-axis is a single snapshot, *i.e.*, 250 fs.

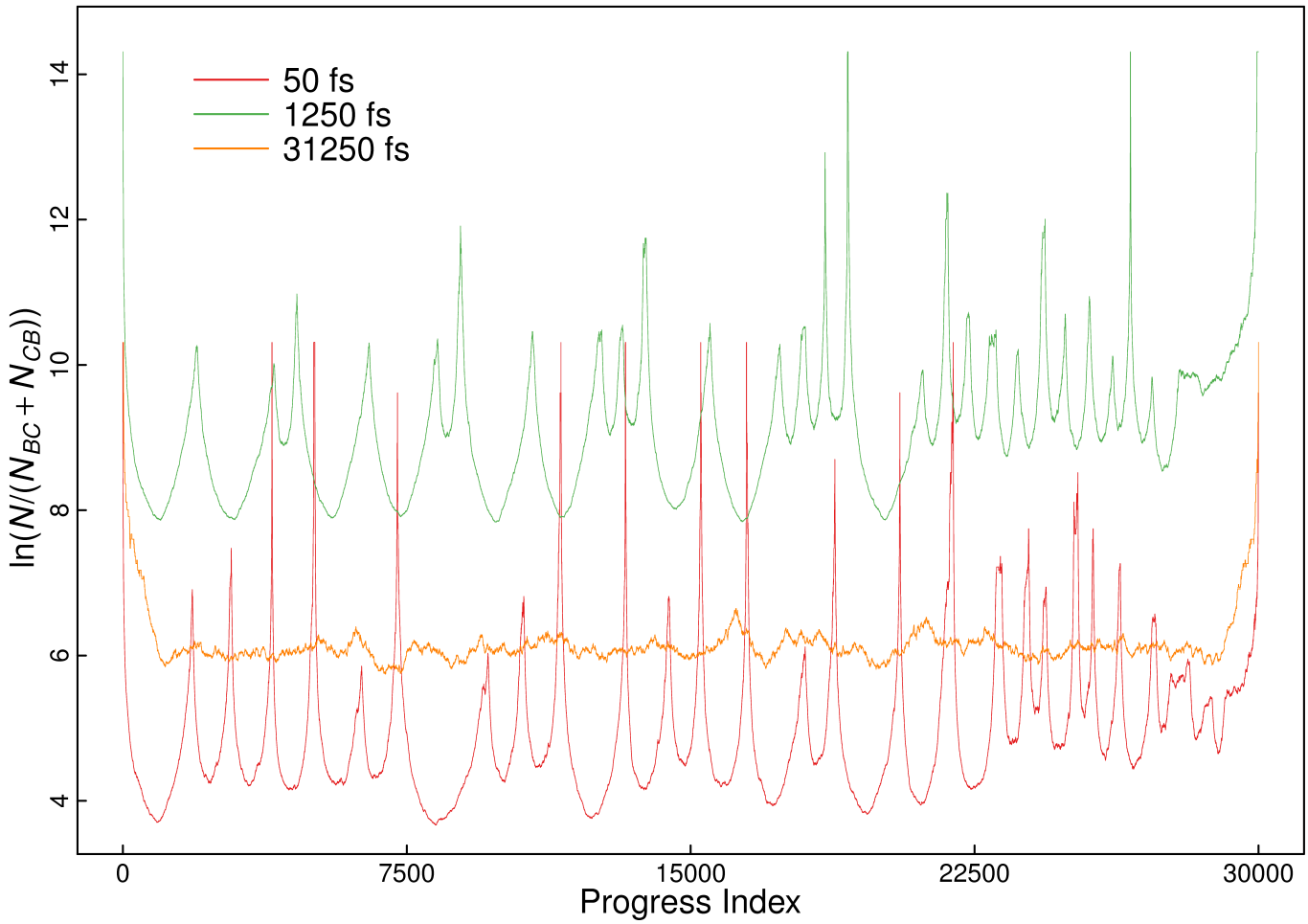


Figure S.2: Influence of temporal resolution on the alternative cut function l with fixed n_l of 1000. This figure is similar to Fig. 3 in the main text, but only three cases are shown. The profile at 1.25 ps is shifted by 4 units for better readability. In analogy to Fig. 3, it can be seen that at 31.25 ps the profile loses its salient features. In contrast to Fig. 3, however, the underlying prior function appears to be flat for all points outside of the first and last n_l points. The cases with finer time resolution therefore highlight two important advantages of annotation function l when compared to function c (see Fig. 3). First, the lack of inherent curvature improves the ability to resolve basins. Second, the localization of the cut improves the signal-to-noise ratio when considering the ratio of values at barriers to those in the bottom of basins. Both advantages are contingent upon finding appropriate values for n_l. For each curve the implied unit of time on the y-axis is a single snapshot of the respective trajectory, *i.e.*, the saving frequency or temporal resolution itself.

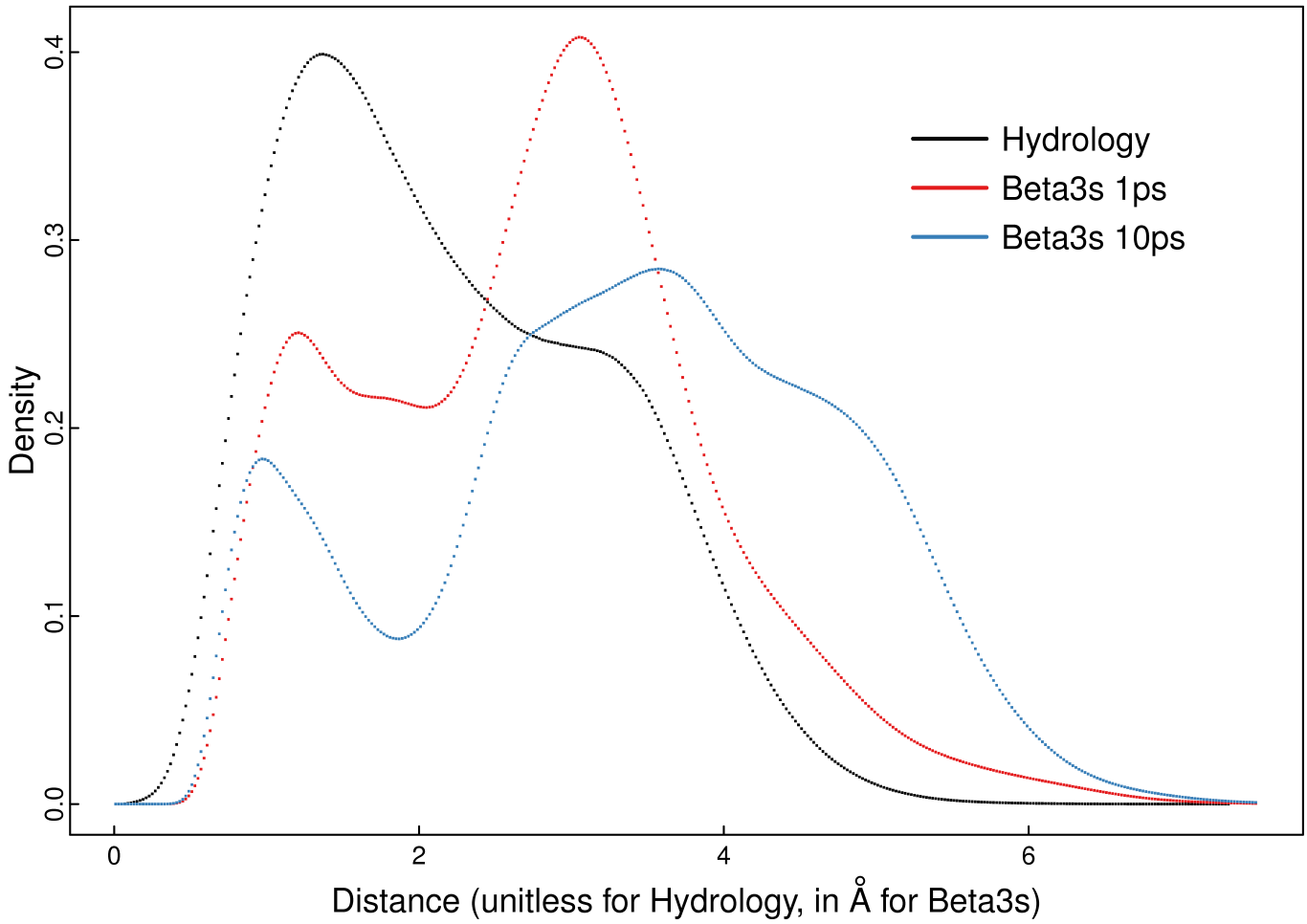


Figure S.3: Histograms of pairwise distance for the hydrology data set in comparison to protein folding data sets. As discussed in Section 3.1 of the main text, the distance spectrum for the hydrology data is expected to be relatively featureless. This is confirmed by the comparison shown here to continuous data obtained for the Beta3S miniprotein [8] at two different time resolutions. In all cases, all possible, unique distance values are computed for $N = 87840$. The figure shows that the hydrology data give rise to a dominant fraction of similar pairs of snapshots. This creates degeneracy in establishing near-neighbor relations that the MST relies on. Even for a total length of just 87.84 ns, the protein data exhibit a much smaller fraction of similar pairs providing more meaningful neighbor relations. This is despite the fact that 87.84 ns are not enough to ensure recurrence *between* major basins. The spectrum becomes increasingly discriminatory if the total time considered is increased (here, up to 878.4 ns).

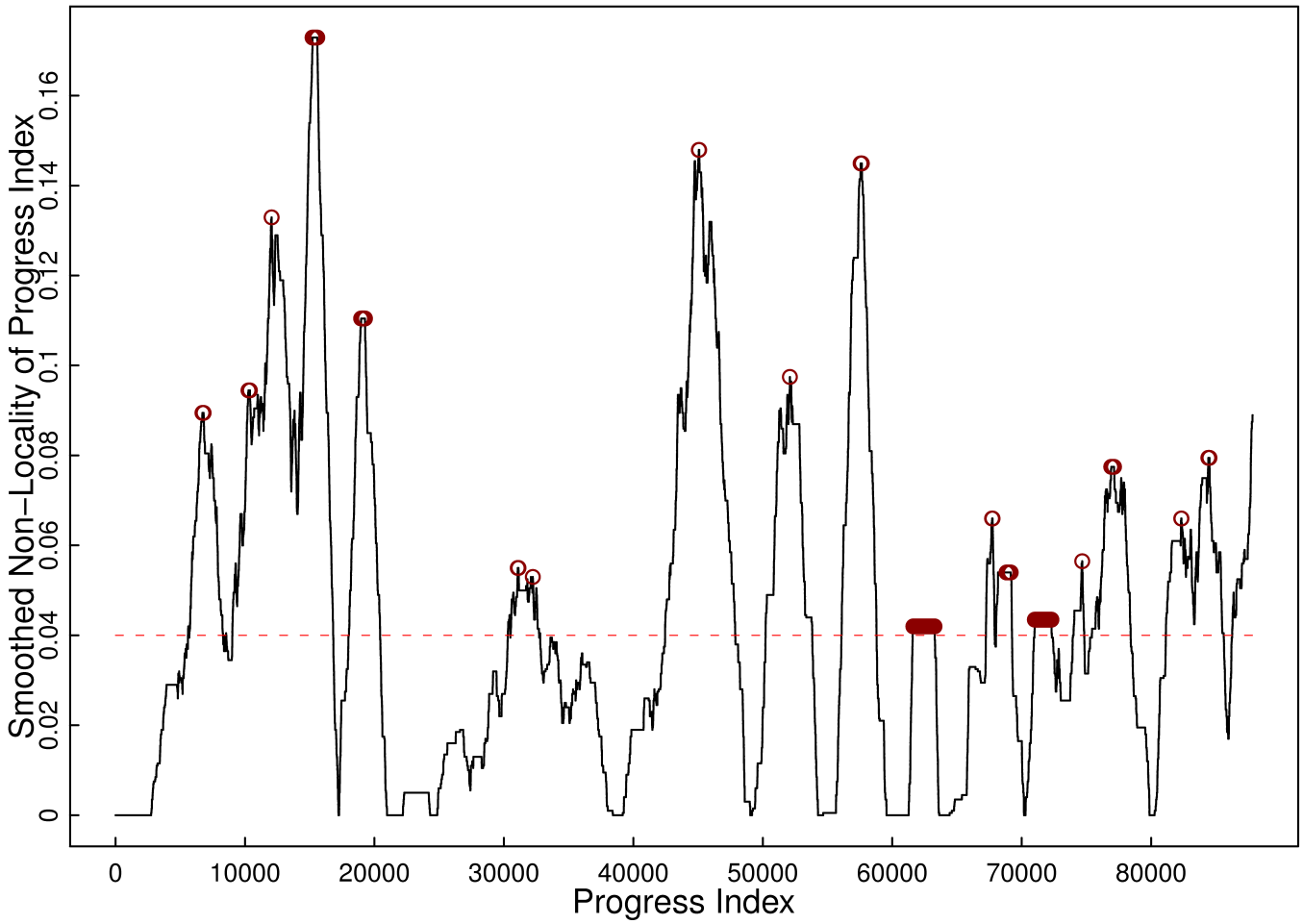


Figure S.4: Identification of barrier regions via non-locality in progress index generation. As discussed in Sections 3.1 of the main text and S.2.1 above, the hydrology data have a poor signal-to-noise ratio, which makes it challenging to identify basins purely based on the annotation functions c or l. During progress index generation, the MST provides information regarding the source or parent vertex (snapshot) that the currently added snapshot is indeed closest to. The value of the progress index for this parent is also known (necessarily smaller), and the difference provides information whether parent and child are likely to be members of the same basin. With a difference threshold of 2000 snapshots, we generate a bit-sequence (1/0) of nonlocality. This bit sequence is then smoothed using (sliding) window averaging with a window size of 2000 snapshots, and the resultant curve is plotted here. Clearly, there are well-defined regions where the function peaks, and we can define a threshold (red dashed line) to select a number of candidate points. It is important to note that the sliding window is not centered at each point, but rather extends only to the left (lower progress index). This is to compensate for the intrinsic property of the progress index in accumulating fringe and transition region points at the right end (toward higher progress index). Due to construction, several points within a sliding window may have the same maximal value. In such a case, we consider only that point yielding the maximal value of the annotation function of interest (c or l).

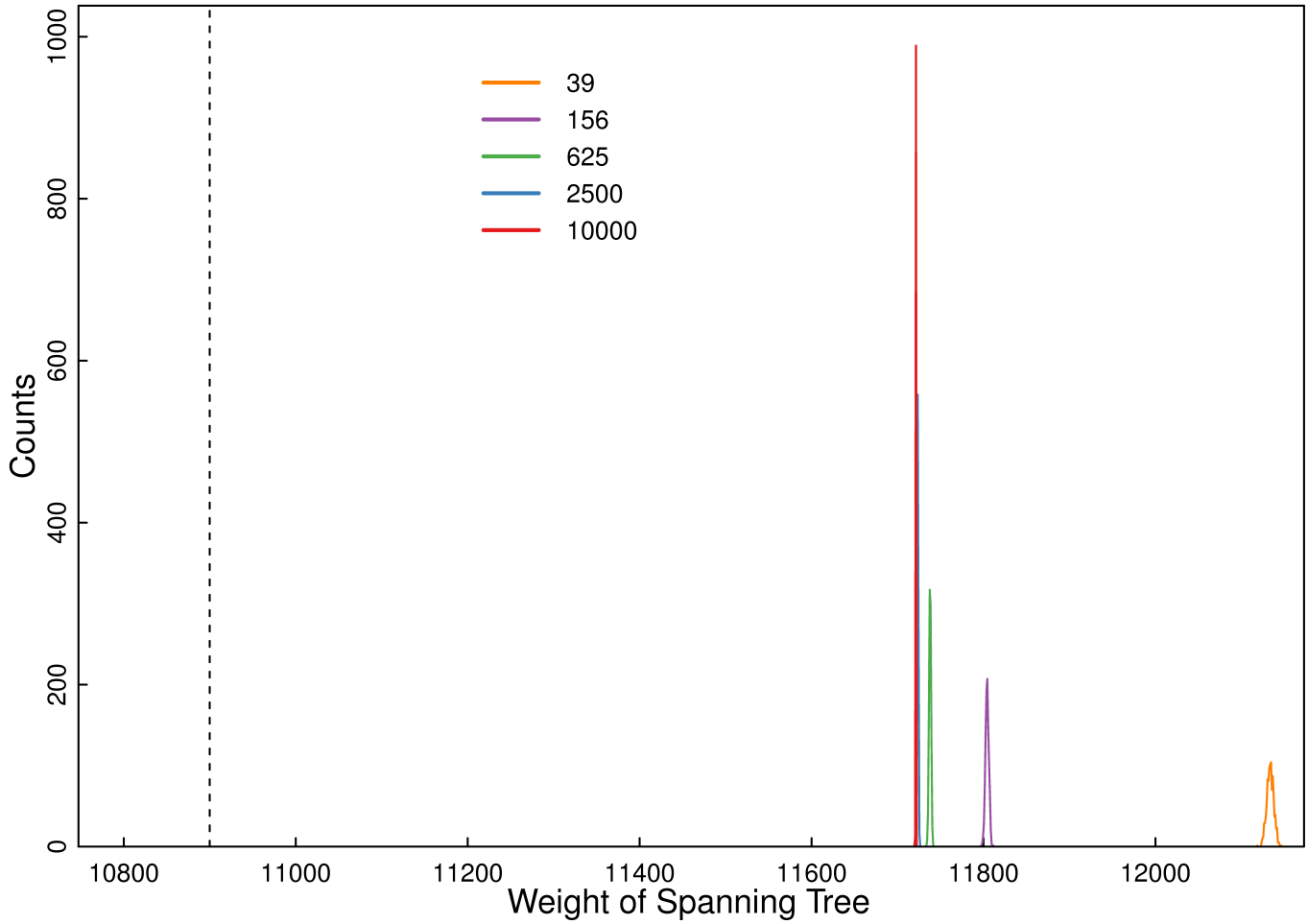


Figure S.5: Quality of the short spanning tree (SST) as a function of the number of guesses. Since the approximate algorithm has a random component and a parameter that is expected to control the total weight of the SST in systematic fashion, this figure shows histograms from 1000 samples each for various choices of N_g. The underlying data are the hydrology data as described in Section S.1.4.3. As expected, larger values for N_g systematically shift the mean of the corresponding histogram toward lower total SST weights. The effect is more pronounced toward low values of N_g with a level of saturation being reached toward high values. The stochasticity of results is also expected to decrease with increasing N_g, and this is evident in the decreasing width of histograms. Lastly, the unique total weight of the MST is indicated as a vertical, dashed line. Clearly, the MST does not appear to coincide with the asymptotic limit for $N_g \longrightarrow \infty$. There are several putative reasons for this, but most likely it is a direct result of the data preorganization exploited during SST construction that limits the search space for a new edge.

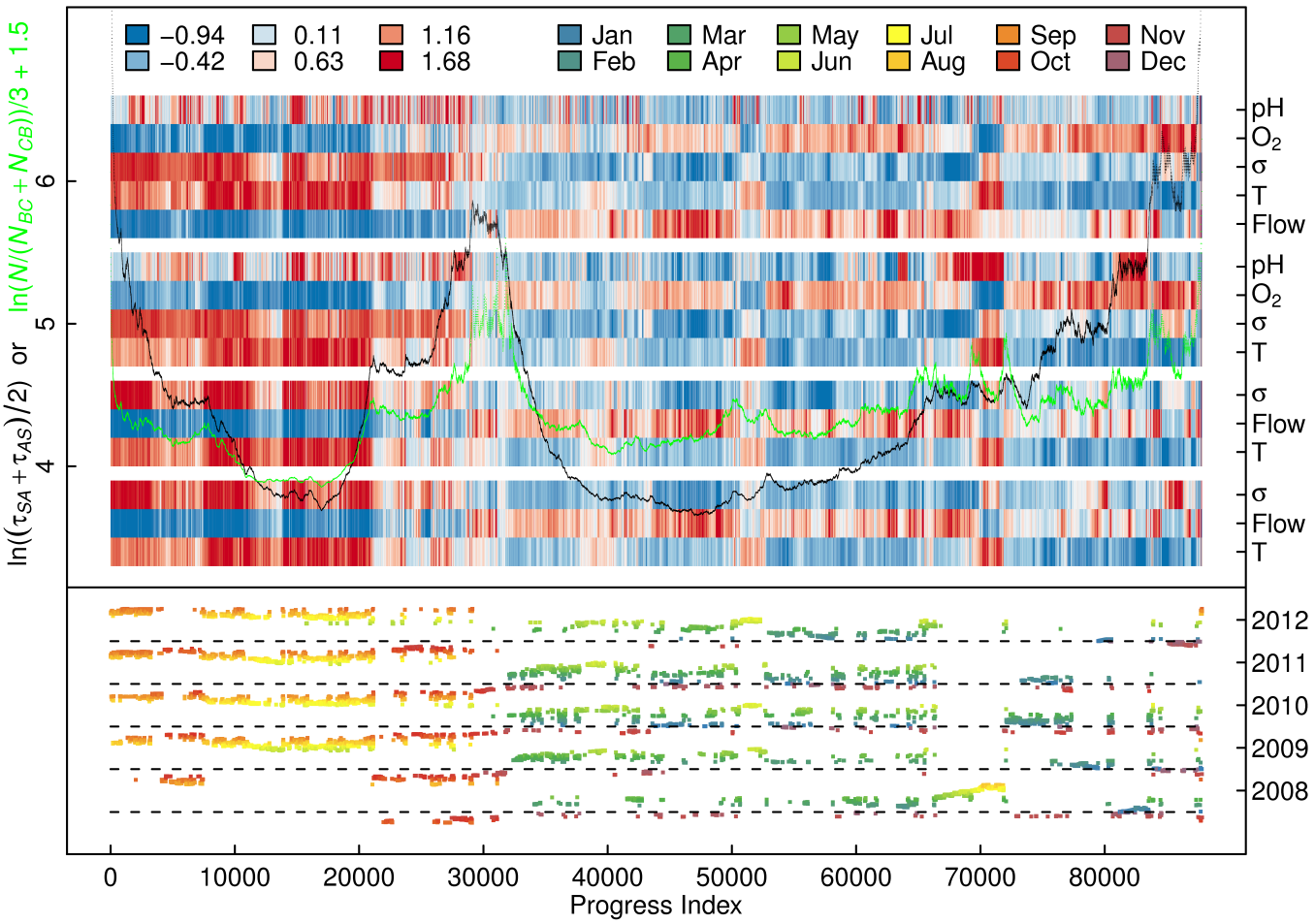


Figure S.6: Comparison of exact and approximate algorithm for the hydrology data set. This figure is analogous to Fig. 6 in the main text, and the reader is referred to the caption of Fig. 6 for understanding what is plotted. Here, we show a progress index generated from the same starting snapshot as in Fig. 6 with the approximate algorithm using a setting of $N_g = 1000$ along with the resultant annotations. The parameters for the tree-based clustering were chosen as 24, 7.0, and 0.36 for the tree height, coarsest threshold, and finest threshold, respectively. The resultant SST had a total weight of 11987 compared to the MST weight of 10906. The plot shows a similar partitioning and arrangement of basins to Fig. 6 with the warm months first followed by a broad basin of the cold season. The remainder of the plot is a series of smaller basins often restricted to individual years. There are two major differences compared to Fig. 6. First, here the transition seasons are more closely grouped and/or are integrated into the large basins leading to a lack of "entropic" regions. Second, the delineation of basins using function l and in particular using function c is much improved. Note for example how close to progress index values of $7 \cdot 10^4$ the spring and mid-summer basins of 2008 are resolved and separated from one another in both annotation functions. This is likely the result of randomness in picking the next snapshot within a group of similar microstates, which creates "artificial" recurrence. It is important to point out that this is neither automatically the case (see winter basin of 2008 just beyond values of the progress index of $8 \cdot 10^4$) nor necessarily desirable. Overall, this figure and Fig. 6 in the main text are similar, however.

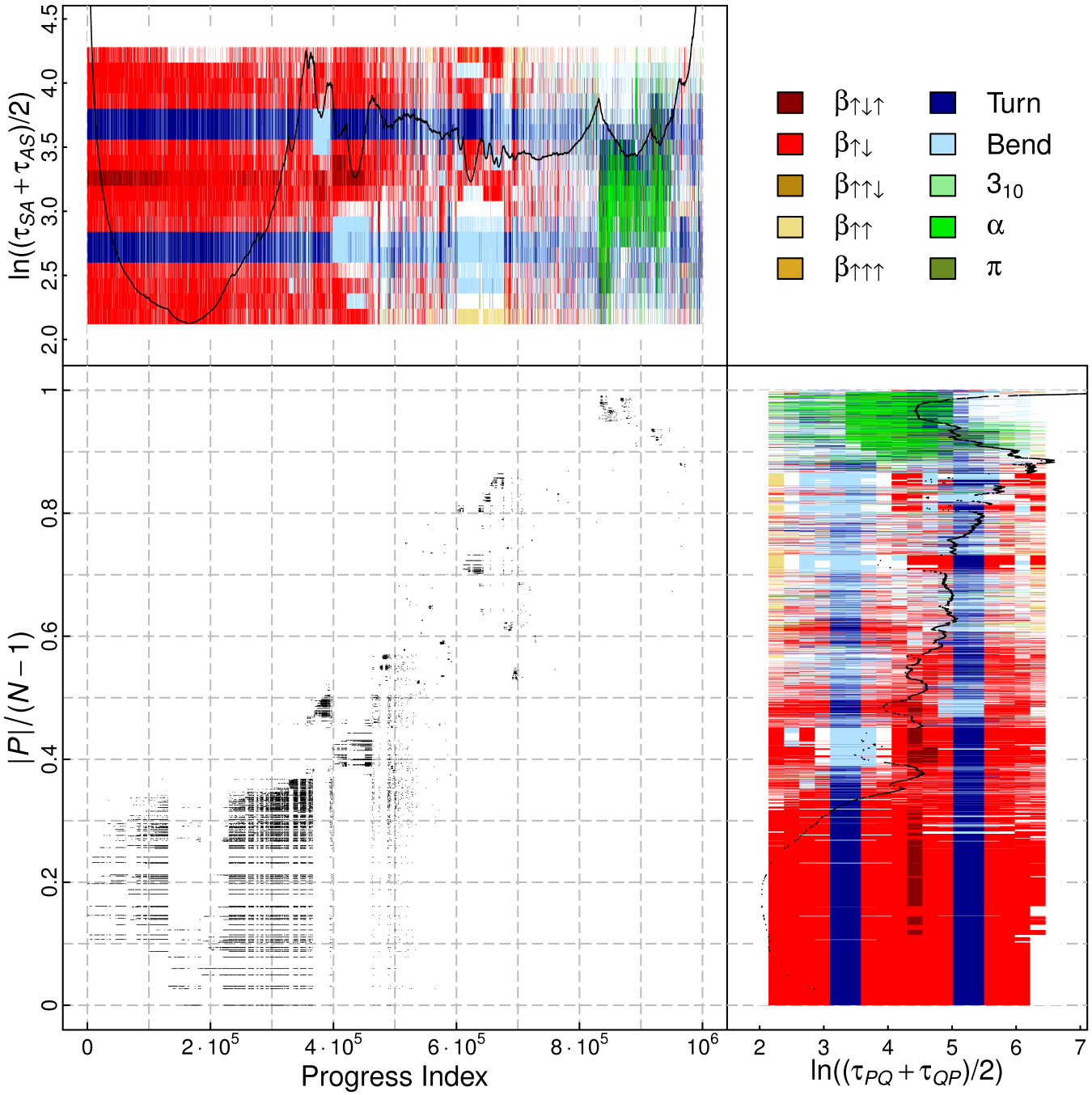


Figure S.7: Comparison of proposed scheme to cut-based free energy profile for Beta3S. The top portion of the plot reproduces the annotated progress index found in Fig. 7 of the main text. Only annotations with function c and DSSP secondary structure assignments [14] are shown for every 20th snapshot. In the DSSP case, this implies that the plotted DSSP strings are extended correspondingly, which means that they are not necessarily exact for 95% of snapshots. At typical resolutions, the similarity is high enough, however, that the visual appearance is not altered by this (this also holds for Figs. 7 and S.8). The right-hand side shows an analogous plot with a cut-based free energy profile [15] instead. Here, the progress index is replaced by a cumulative probability ($|P|/(N-1)$) computed from the equilibrium probabilities of all the nodes of the underlying conformational network that are part of the growing set P. The ordering principle is kinetic distance from a reference state (here, the native basin). A network cut is used to annotate this sequence of kinetically ordered nodes, and the similarity is apparent. To reduce image size, annotations are only shown for those clusters with at least 10 members (they encompass about 74% of snapshots). For DSSP color annotations, this means specifically that the plotted DSSP string of the centroid of a given cluster extends until the integrated weight reaches a new cluster of a minimum size of 10. Because the omitted clusters are tiny, the visual appearance of the plot is not altered at typical resolutions. The lower left portion plot shows a scatter plot emphasizing the correspondence between network nodes of size 100 or

larger and the positions of their constituent snapshots in the progress index. By means of the DSSP annotations, it is easily seen that the same basins are resolved with the same widths (total weights). This suggests that the proposed algorithm is unlikely to suffer from false positives or false negatives in terms of partitioning the data into basins if the underlying trajectory is sufficiently recurrent. Additionally, clusters of points are largely close to the diagonal indicating good correlation between the explicit, kinetic ordering and the progress index for this system. The implied unit of time for the mean first passage times is one snapshot, *i.e.*, 20 ps.

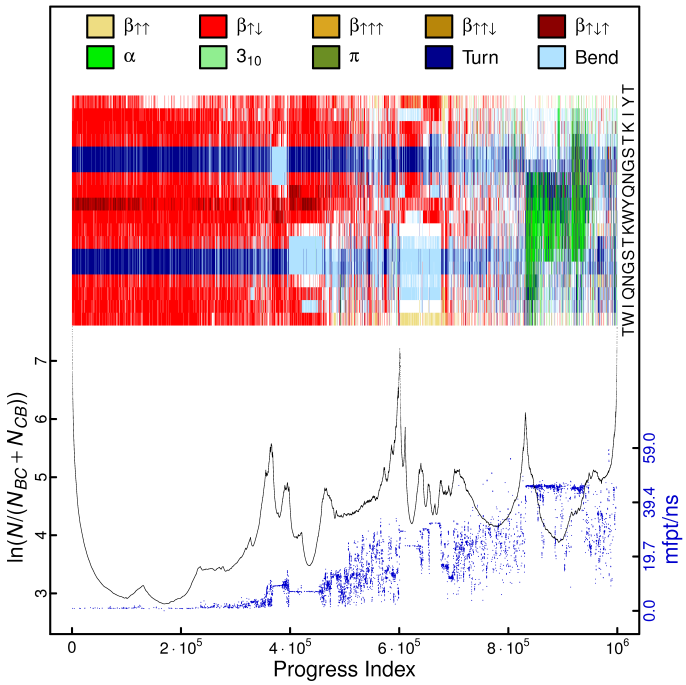


Figure S.8: Results for Beta3S with the alternative cut function l and fixed n_l of 10^5. This figure is identical to Fig. 7 in the main text with the exception that the primary annotation function is not c, but rather l (most importantly, it uses the exact same SST). Because of the wide distribution of sizes of basins for this system, a flat choice for n_l is not expected to enhance resolution uniformly. Interesting changes compared to Fig. 7 involve a pronounced barrier at $6 \cdot 10^5$ and substructure within the native basin, the latter of which is more apparent here, but also discernible in Fig. 7. While not clearly visible in the plot, the data allude to the fact that, for small enough and fixed n_l, the number of direct transitions between partitions B and C starts to *decrease* for increasing basin size. This leads to a flattening effect that makes it difficult distinguish large basins from "entropic" regions, and this effect is much more apparent for $n_l = 2 \cdot 10^4$ (not shown).

References

[1] J. B. Kruskal Jr., On the shortest spanning subtree of a graph and the traveling salesman problem, Proc. Am. Math. Soc. 7 (1956) 48–50.

[2] J. Nešetřil, E. Milková, H. Nešetřilová, Otakar Borůvka on minimum spanning tree problem, Translation of both the 1926 papers, comments, history, Discrete Math. 233 (2001) 3–36.

[3] A. Vitalis, A. Caflisch, Efficient construction of mesostate networks from molecular dynamics trajectories, J. Chem. Theory Comput. 8 (2012) 1108–1120.

[4] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, J. Am. Chem. Soc. 118 (1996) 11225–11236.

[5] P. Ferrara, J. Apostolakis, A. Caflisch, Evaluation of a fast implicit solvent model for molecular dynamics simulations, Proteins: Struct., Funct., Bioinf. 46 (2002) 24–33.

[6] S. Muff, A. Caflisch, Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein, Proteins: Struct., Funct., Bioinf. 70 (2008) 1185–1195.

[7] S. V. Krivov, S. Muff, A. Caflisch, M. Karplus, One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process, J. Phys. Chem. B 112 (2008) 8701–8714.

[8] T. Zhou, A. Caflisch, Free energy guided sampling, J. Chem. Theory Comput. 8 (2012) 2134–2140.

[9] B. Qi, S. Muff, A. Caflisch, A. R. Dinner, Extracting physically intuitive reaction coordinates from transition networks of a β-sheet miniprotein, J. Phys. Chem. B 114 (2010) 6979–6989.

[10] U.S. Geological Survey, Site inventory for the nation, http://waterdata.usgs.gov/nwis/inventory, accessed October 26, 2012.

[11] M. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World map of the Köppen-Geiger climate classification updated, Meteorol. Z. 15 (2006) 259–263.

[12] K. D. Carpenter, S. Sobieszczyk, A. J. Arnsberg, F. A. Rinella, Pesticide occurrence and distribution in the lower Clackamas river basin, Oregon, 2000–2005, Scientific Investigations Report 2008-5027, U.S. Department of the Interior – U.S. Geological Survey, 2008.

[13] WeatherSpark, Historical weather for 2008 in Portland, Oregon, USA, http://weatherspark.com/history/30477/2008/Portland-Oregon-United-States, accessed October 29, 2012.

[14] W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[15] S. V. Krivov, M. Karplus, One-dimensional free-energy profiles of complex systems: Progress variables that perserve the barriers, J. Phys. Chem. B 110 (2006) 12689–12698.

Chapter 3

High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations

Blöchliger, N., Vitalis, A. and Caflisch, A. *Scientific Reports*, 4: 6264, 2014

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PROTEINS
COMPUTATIONAL BIOPHYSICS
COMPUTATIONAL SCIENCE

Received
25 April 2014

Accepted
15 August 2014

Published
2 September 2014

Correspondence and requests for materials should be addressed to A.V. (a.vitalis@bioc.uzh.ch) or A.C. (caflisch@bioc.uzh.ch)

High-Resolution Visualisation of the States and Pathways Sampled in Molecular Dynamics Simulations

Nicolas Blöchliger, Andreas Vitalis & Amedeo Caflisch

University of Zurich, Department of Biochemistry, Winterthurerstrasse 190, CH-8057 Zurich.

We have recently developed a scalable algorithm for ordering the instantaneous observations of a dynamical system evolving continuously in time. Here, we apply the method to long molecular dynamics trajectories. The procedure requires only a pairwise, geometrical distance as input. Suitable annotations of both structural and kinetic nature reveal the free energy basins visited by biomolecules. The profile is supplemented by a trace of the temporal evolution of the system highlighting the sequence of events. We demonstrate that the resultant SAPPHIRE (States And Pathways Projected with HIgh REsolution) plots provide a comprehensive picture of the thermodynamics and kinetics of complex, molecular systems exhibiting dynamics covering a range of time and length scales. Information on pathways connecting states and the level of recurrence are quickly inferred from the visualisation. The considerable advantages of our approach are speed and resolution: the SAPPHIRE plot is scalable to very large data sets and represents every single snapshot. This minimizes the risk of missing states because of overlap or prior coarse-graining of the data.

Present day science or more broadly society record observations as a function of time in diverse contexts[1]. Data on meteorological phenomena, communication tracking, or financial markets, to name a few, are all mined for the generation of predictive models. The raw data are usually unfit for human consumption due to their high dimensionality and sheer size[2,3]. Both aspects limit the types of analyses performed on these "big data" to those algorithms with satisfactory scaling properties[4]. In biophysics, long computer simulations of the trajectories of complex macromolecules with high-dimensional representations have become commonplace[5], and this is where our particular interest lies[6].

Molecular dynamics (MD) simulations of proteins and other biomolecules[7] record stochastic trajectories, in which the macromolecule visits a number of different, metastable states (free energy basins) connected by an ensemble of pathways of interconversion. The latter report on the barriers of the underlying free energy landscape[8]. Because millions of snapshots are now routinely recorded for thousands of coupled degrees of freedom[3], MD trajectories call for scalable algorithms that are able to provide information-preserving projections for this specific class of complex systems. We have recently introduced such an algorithm[9] and provide a brief description next.

Given a definition of distance between trajectory snapshots, the entire data set is considered as a complete graph with vertices corresponding to snapshots and edge weights given by the pairwise distances between snapshots. Either the exact or an approximation to the minimum spanning tree are computed. From a generally arbitrary starting point, the available edge with smallest weight is followed to define a sequence of snapshots, the so-called progress index. The available edges at each point are those connecting any snapshot not considered yet (we refer to this set of snapshots as A) with any snapshot already included (the set S). The resulting sequence has the crucial property of stepping through high density regions one by one. It can therefore be expected that all free energy basins will appear as groups of nearby points along the progress index. Importantly, the progress index does not reflect the temporal nature of the input data in any way, *i.e.*, it is generally independent of input order. Because every snapshot is considered, the limiting resolution is optimal given the time resolution of the input trajectory.

The progress index can be annotated both kinetically and structurally to provide an informative and compact representation of all major states visited by the input trajectory. The procedure has several advantages over projections using geometric or kinetic distances from a reference state to order snapshots. First, it maximizes resolution as mentioned above. Second, it avoids overlap precisely because the ordering is not with respect to a

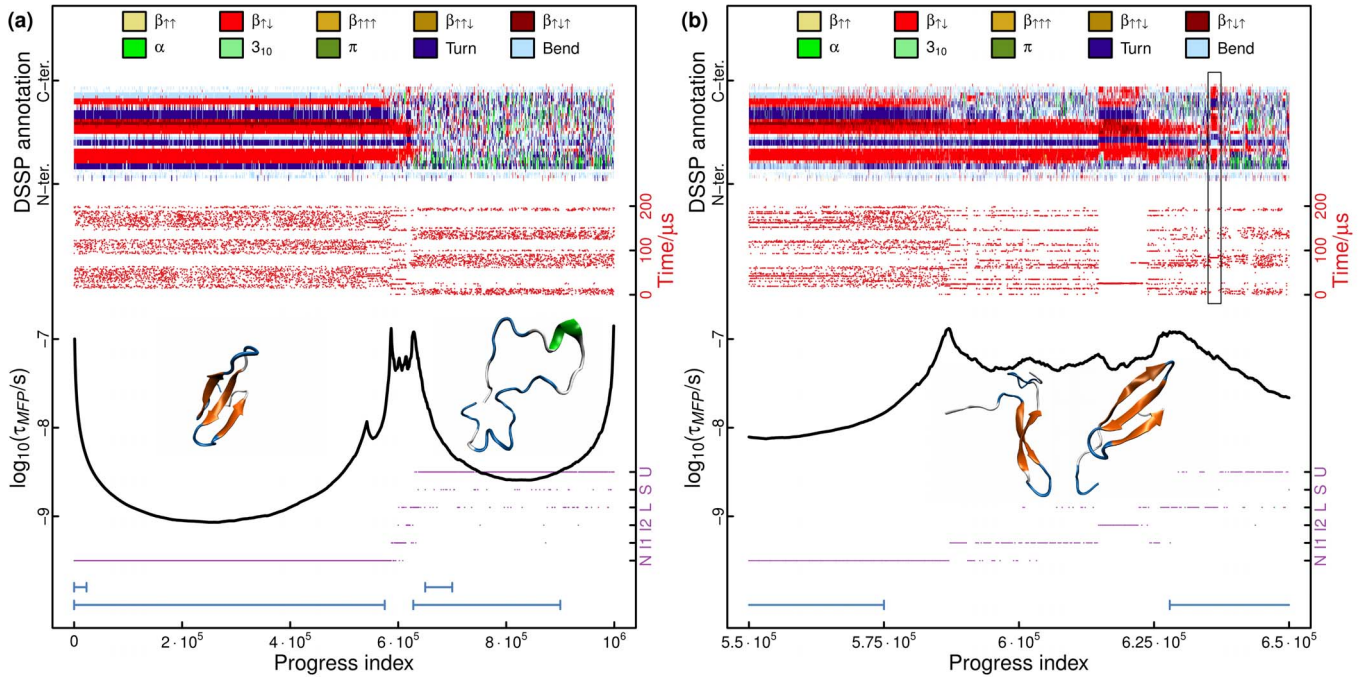


Figure 1 | SAPPHIRE plot for FiP35. (a) The progress index, of 10^6 snapshots from 200 μs of MD data, is annotated with kinetic information (τ_{MFP}, black curve), dynamical trace (red dots), DSSP assignment[17] by residue (legend on top) and the state partitioning of Berezovska *et al.*[20] These annotations are only shown for every 1000th, 100th, 1000th and 500th snapshots, respectively, in order to maintain readability at fixed figure resolution. The limits of possible definitions of the folded and unfolded states for the computation of transition path times are indicated by the blue, horizontal lines. Cartoons[31] of a snapshot in the native state and an unfolded conformation are shown. (b) Zoom-in on the transition region of the SAPPHIRE plot shown in (a). The various annotations are shown for every 100th, 10th, 50th, and 250th snapshots, respectively. Representative conformations of I1 and I2 are shown as cartoons. The box highlights a particular state (see text).

particular state. Third, it is easy to use, scalable, and requires a notion of distance between snapshots as the only "parameter". Like most data mining methods exploiting pairwise similarity as a guide, *e.g.*, clustering[10], it requires sufficient sampling density. The sampling weights of individual basins can in general be resolved quantitatively.

In the present, short contribution, we apply the method of Blöchliger *et al.*[9] to two molecular dynamics trajectories of proteins, which were produced using dedicated hardware[11]. We annotate the plot threefold, *viz.*, structurally, kinetically, and with times of occurrence in the original trajectories (called dynamical trace hereafter). We demonstrate that the information summarized in the resultant SAPPHIRE (**S**tates **A**nd **P**athways **P**rojected with **HI**gh **RE**solution) plot provides an efficient means of identifying the statistically reliable states visited by a complex, dynamical system while enabling a rapid assessment of state interconversion and recurrence, which provide information on kinetic pathways and simulation convergence, respectively.

Results

We present results on two different proteins. The data on Fip35[12], a small WW domain, come from two long MD trajectories and describe reversible transitions of this peptide between the folded state, a three-stranded β-sheet, and a coil-like unfolded state. The single MD trajectory obtained for the 58-residue bovine pancreatic trypsin inhibitor (BPTI), a protein with a mixed α/β fold[13], exhibits few transitions between distinct folded states that differ prominently in the isomerization states of disulphide bridges. We analyse both of these data sets with SAPPHIRE plots. The general annotation functions we use are as follows:

1. The sets *A* and *S* allow us to stipulate a two-state Markov state model, and we can derive the mean first passage times in either direction[9]. A cut function as used elsewhere[14] allows an analytical evaluation. We define the average of the two values as τ_{MFP}. This kinetic annotation is expected to highlight barriers reliably with the caveat that it cannot be interpreted quantitatively due to the simplicity of the two-state model. For data sets obtained by concatenating many short MD trajectories, the cut function is adjusted to ignore the spurious transitions at the break points between two trajectories.
2. The actual time of occurrence in the input data (dynamical trace) is plotted for each snapshot as an annotation highlighting direct transitions between states (pathways). Because the progress index is expected to be free of overlap, this allows a straightforward assessment of recurrence. This annotation is less informative if the data set is a concatenation of short trajectories where each continuous segment visits only one or few basins.
3. States themselves are characterized by a structural annotation. This is necessarily system-specific and requires prior knowledge of the system and data. An informative, geometric annotation can be exceptionally helpful in connecting the states identified by the kinetic annotation with a structural interpretation fit for human consumption. Structural annotations do not depend on input order, *i.e.*, they are useful even for unordered input data.

Reversible folding of a 35-residue protein domain. FiP35[12] exhibits reversible folding at a simulation temperature of 395 K in explicit solvent molecular dynamics runs of a total length of 200 μs. Specifically, the trajectories show that FiP35 converts 10–15 times between an unfolded state that is very low in secondary structure content and the native topology, *viz.*, a twisted, three-stranded β-sheet[11,15,16]. All following results refer to a specific computational model and sampling protocol[11] underlying the trajectories being analysed. Due to the protein's small size, it is possible to provide a comprehensive, structural representation at the backbone level using a DSSP annotation[17] resolved by residue. Fig. 1(a) shows the

SAPPHIRE plot for the composite trajectory using this annotation, and it is immediately apparent that the native topology is observed more than 50% of the time. The native basin is delineated by the kinetic annotation (black line) as expected. The unfolded state shows no consistent secondary structure and is kinetically homogeneous, suggesting that FiP35 should be described well as a two-state folder.

Fig. 1(b) highlights one of the major advantages of our approach, *i.e.*, its high resolution. Here, we zoom in on the transition region. Previous analyses of the same data suggested the existence of at least two intermediates[18–20], and we have additionally annotated the SAPPHIRE plot with the state partitioning proposed by Berezovska *et al.*[20] Referring to their Fig. 3, we denote the larger and smaller of the two unlabelled states as L and S, respectively. By also taking into account the dynamical trace, Fig. 1 allows us to quickly extract the following results:

- I2 is identified with a very homogeneous state sampled extensively only once during the 200 μs and characterized by a three-stranded β-sheet topology with shifted registry for the N-terminal hairpin (see DSSP annotation in Fig. 1(b)). It is referred to as a kinetic trap elsewhere[16], and its sampling weight is 0.5–1.0%.
- In state I1, only the N-terminal hairpin is formed with the C-terminus largely coil-like. This is the state sampled most often when converting between folded and unfolded states (F and U, respectively) via an intermediate, and its weight is ∼2.5%.
- Structurally, L consists largely of a boundary (barrier) region between I1 and U. Our data suggest that the kinetically and geometrically homogeneous state highlighted by the black box in Fig. 1(b) should have been separated out.
- States U and F are explored for extended periods of time less than 10 times each. Several excursions into intermediate states are unproductive.
- We cannot assign obvious meaning to state S.
- Fig. 1(a) suggests the existence of an additional state with a weight of ∼5%, in which the N-terminal turn is in an alternative conformation. Similar differences are observed when comparing NMR structures of *apo* and *holo* forms of related WW domains[21]. While kinetically distinct, this state may have been ignored by Berezovska *et al.* because it is not on-pathway. Indeed, it is likely to correspond to the state labelled *holo* by Lane *et al.*[19].

The picture emerging from the above is overall congruent with analyses of the same data reported elsewhere[16,18–20,22,23]. This also extends to the pathway information regarding dominant routes of folding. The main advantage of carefully constructed Markov state models is of course that probability flux and time scales are explicit and quantitative. This bears the caveat that the required lag times may be so large that useful information on or below the timescale of this lag time is lost. Conversely, Fig. 1 allows many of the important conclusions arrived at in the literature from a single plot that can be produced in near-linear time with respect to the number of snapshots. Some of the kinetic information such as probability flux and pathways are qualitative in nature only. It may be necessary to rescale the plot to resolve some of the finer details. It is also important to keep in mind that the sequence from left to right does not correspond to real pathways taken by the trajectory even though it may sometimes appear that way. Pathway information is gleaned exclusively from the dynamical trace, which is read vertically from bottom to top.

The SAPPHIRE plot allows a very straightforward grouping of snapshots into states. From these groupings, we can compute further quantities such as the times taken to reach the folded state from the unfolded state and vice versa (transition path times). Fig. 1(a) indicates two extreme definitions of folded and unfolded states, and within these limits the computed transition path times range from 20 to 180 ns. Experiments suggest more roughness of the underlying landscape leading to longer transition times (∼1 μs)[24]. We reemphasize that Fig. 1 is conditional upon a specific model, *i.e.*, force field, and sampling protocol, which in all likelihood are prone to both systematic and statistical errors. We merely perform the analysis here to demonstrate how the SAPPHIRE plot is also an excellent starting point for further efforts in characterizing and understanding the data and system at hand.

Native state dynamics of a folded protein. Analyses of MD simulations of the folded state ensemble of the 58-residue bovine pancreatic trypsin inhibitor (BPTI) have revealed that a number of states identified by NMR experiments[25,26] are populated significantly in the trajectories albeit with inaccurate weights. These metastable states can often be correlated with isomerizations of the disulphide bridges, in particular Cys14-Cys38[27,28]. Shaw *et al.*[11] used a stochastic algorithm to obtain a coarse, kinetic clustering of their 1.03 ms MD trajectory of BPTI relying on the autocorrelation of interatomic distances. Empirically, they found that five states with significant populations could be identified reliably. These states were annotated structurally. The most important states (the smaller of the two is the one most resembling the crystal structure) are clearly seen in the SAPPHIRE plot as the first two basins from the left (Fig. 2). The structural annotation we select here confirms that the barrier identified by the kinetic analysis (black line) is related to the isomerization of the disulphide bond Cys14-Cys38. The dynamical trace uses the colour scheme of Shaw *et al.* (distinguishing the red, blue, green, purple, and black states). It unmasks that both major states are long-lived and that there is a clear separation of time scales with respect to the mixing time within each basin.

Fig. 2 indicates that there is a mismatch in assignment between that of Shaw *et al.* and the positions on the x-axis for several, short excursions into a given state. As an example, we consider the highlighted trajectory segment sampled at ∼0.5 ms that is annotated by Shaw *et al.* to be in the red state but that is placed by the SAPPHIRE plot in the basin corresponding to the blue state. To understand why this may be the case, we first note that the structural annotations generally reveal a small amount of mixing that may be considered erroneous. Indeed, for the segment in question, inspection of instantaneous values yields that the Cys14 side chain angles adopt the values for the blue state, but the χ_3 angle and the χ_2 angle of Cys38 do not (not shown). The combination of values for the dihedral angles places this segment outside of the list of states characterized previously[28]. It appears kinetically homogeneous and may correspond to an incomplete or blocked transition. Its sampling weight is so low that neither the SAPPHIRE plot nor the kinetic clustering are sensitive enough to resolve it as an independent state. Due to its intermediate nature, it is lumped into either one of the adjacent states. A very similar effect is observed for a second, highlighted segment (at ∼0.75 ms), for which just the two Cys38 side chain angles deviate from the blue state.

The SAPPHIRE plot for BPTI also reveals that over the course of the 1.03 ms trajectory the purple and black states are sampled extensively just once and twice, respectively. This allows us to infer a lack of recurrence, *i.e.*, sampling weights are unlikely to be converged. Poor sampling may also limit the number of states obtainable from Markov state models[29] and decrease the accuracy of any extracted passage times. The bottom panel of Fig. 2 zooms into a very thin time slice to illustrate the pathway taken to reach the black state. This is annotated by cartoons and a specific, interatomic distance involving a residue identified by the original authors as being discriminative for this state[11]. The final result we want to mention in this short note is that the SAPPHIRE plot suggests the green state to be partitioned further. The kinetic annotation is consistent with the dynamical trace in that the two major substates of the green state are homogeneous with respect to the times they were sampled at (no mixing). This is

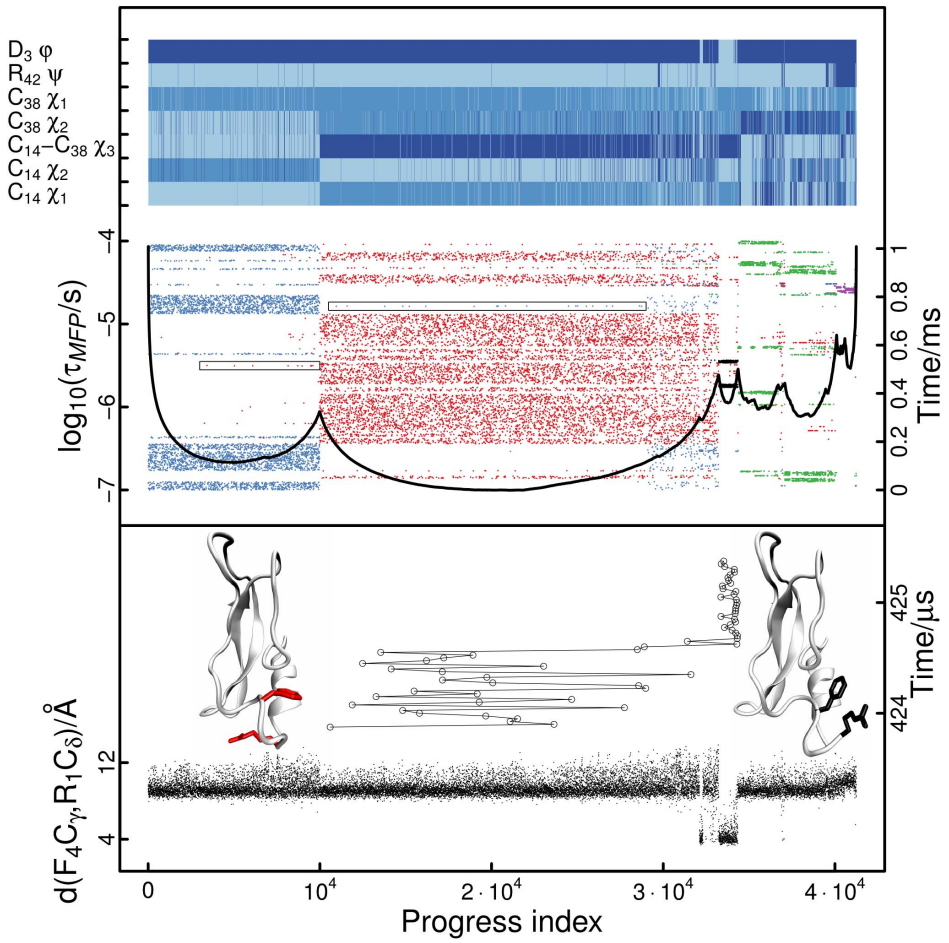


Figure 2 | SAPPHIRE plot for BPTI. (Upper panel) The progress index, of 41250 snapshots from 1.03 ms of MD data, is annotated with kinetic information (τ_{MFP}, black curve), dynamical trace (dots coloured according to the kinetic clustering of Shaw *et al.*)[11], and selected dihedral angles. These annotations are only shown for every 20th, 2nd and 2nd snapshots, respectively, in order to maintain readability at fixed figure resolution. The annotation with dihedral angles uses binning into up to three bins with boundaries chosen as follows: Cys14 χ_1 (−120°, −5°, 120°), Cys14 χ_2 (−140°, 0°, 130°), Cys14-Cys38 χ_3 (0°, 150°), Cys38 χ_2 (−155°, −105°, 120°), Cys38 χ_1 (−120°, 0°, 140°), Arg42 ψ (−100°, 75°), and Asp3 φ (0°, 100°). These boundaries were obtained from direct inspection of the individual histograms for each angle. Boxes highlight two brief stretches of the trajectory referred to in the text. (Lower panel) Zoom-in on a thin time slice of the dynamical trace to visualise a particular transition from the red to the black state. End points of this transition are shown as cartoons with Arg1 and Phe4 in a stick-like representation[31]. The plot is annotated further by the distance between the C_γ atom of Phe4 and the C_δ of Arg1, which is shown for every 5th snapshot.

despite the fact that they appear to be directly adjacent to one another in terms of transition pathways.

We conclude the description of the performance of the SAPPHIRE plot with a note of caution. In Fig. 2, toward the right side of the largest basin, there is a region of both temporal and geometric ambiguity most clearly seen by the overlap of blue and red dots in the dynamical trace. Here, the progress index is placing "fringe" regions of *both* basins. This weakness results from an insufficient sampling density for these lower likelihood regions that immediately surround well-defined states. It is rectified by having better time resolution or, at the risk of a decrease in resolution, by lowering the dimensionality of representation. We show the data on the sparsely sampled trajectory here to illustrate both the general robustness and possible errors encountered with smaller data sets.

Discussion

With growing computing resources and growing data sets, it has become paramount to use tools that quickly and efficiently improve our understanding of a system as complex as a biomolecule. The data required for the SAPPHIRE plot with all three annotations can usually be computed in near-linear time in a single run by the CAMPARI simulation and analysis package (http://campari.sourceforge.net). The plots are ideally generated as fully scalable vector graphics. At fixed resolution, readability may be improved by displaying annotations more sparsely, and we have done this for both figures. The required user input is the definition of a suitable measure of pairwise distance, and this choice may also help determine which structural annotations to use.

In Figs. 1 and 2, we have shown that SAPPHIRE plots offer an efficient procedure for the analysis and comprehensive pictorial description of complex systems undergoing stochastic evolution, such as proteins. Thermodynamics are resolved quantitatively, and the construction of the ordering of snapshots minimizes the risk of state overlap. Major basins are delineated easily by all three annotation functions. Qualitative information about pathways is available at the temporal resolution offered by the trajectory itself. The rapid availability of this information is not only valuable *per se* but can also be used to guide further simulations and analyses.

Methods

The algorithm underlying the SAPPHIRE plot has been describe qualitatively above (see Introduction and Results). For a complete description we refer the reader to the original publication[9]. In terms of efficiency, the overall annotation procedure requires

linear time with respect to the number of snapshots. The calculation of the required spanning tree is the most expensive step of the algorithm and is aided by heuristics in either variant (exact or approximate). The approximate version can be scaled to very large data sets. When using this version, it will generally be useful to rerun the analysis a few times due to the stochastic nature of the spanning tree. In particular, the kinetic annotation function is sensitive to where a basin appears in the progress index and how well basins to the left have been captured.

The FiP35 trajectory encompasses 10^6 snapshots saved every 200 ps, while the 41250 snapshots of data on BPTI have a coarser time resolution of 25 ns. Pairwise distances were defined as the coordinate root mean square deviation (RMSD) computed over the backbone oxygen and nitrogen atoms of residues 7–29 for FiP35 and over 695 nonsymmetric atoms for BPTI. These choices reflect the different levels of variance in the two data sets. The approximate algorithm was used for both systems. It requires additional parameters as follows. The number of guesses to find putative nearest neighbours from within a limited space defined by preorganization of the data via clustering[30] was set to a value of 1000 throughout. The lower threshold radii for clusters were 3.0 and 2.5Å for FiP35 and BPTI, respectively, and the upper threshold radii were 10.0 and 3.0Å. The required input data took ~11 and ~5 hours to compute on a single Intel Xeon core (either E5435 or E5410) for Figs. 1 and 2, respectively.

1. Fu, T.-C. A review on time series data mining. *Eng. Appl. Artif. Intel.* **24**, 164–181 (2011).
2. Kehrer, J. & Hauser, H. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Trans. Vis. Comput. Graph.* **19**, 495–513 (2013).
3. Rysavy, S. J., Bromley, D. & Daggett, V. DIVE: A graph-based visual-analytics framework for big data. *IEEE Comput. Graph. Appl.* **34**, 26–37 (2014).
4. Bohlouli, M. *et al.* in *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives* (ed Fathi, M.) Ch. Towards an integrated platform for big data analysis, 47–56 (Springer, 2013).
5. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophys.* **41**, 429–452 (2012).
6. Spiliotopoulos, D. & Caflisch, A. Molecular dynamics simulations of bromodomains reveal binding-site flexibility and multiple binding modes of the natural ligand acetyl-lysine. *Isr. J. Chem.* ***in press***, DOI: 10.1002/ijch.201400009 (2014).
7. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
8. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
9. Blöchliger, N., Vitalis, A. & Caflisch, A. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comp. Phys. Comm.* **184**, 2446–2453 (2013).
10. Xu, R. & Wunsch II, D. C. Clustering algorithms in biomedical research: A review. *IEEE Rev. Biomed. Eng.* **3**, 120–154 (2010).
11. Shaw, D. E. *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
12. Liu, F. *et al.* An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. USA* **105**, 2369–2374 (2008).
13. Berndt, K. D., Güntert, P., Orbons, L. P. M. & Wüthrich, K. Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures. *J. Mol. Biol.* **227**, 757–775 (1992).
14. Krivov, S. V. & Karplus, M. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* **110**, 12689–12698 (2006).
15. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
16. Kellogg, E. H., Lange, O. F. & Baker, D. Evaluation and optimization of discrete state models of protein folding. *J. Phys. Chem. B* **116**, 11405–11413 (2012).
17. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
18. Krivov, S. V. The free energy landscape analysis of protein (FIP35) folding dynamics. *J. Phys. Chem. B* **115**, 12315–12324 (2011).
19. Lane, T. J., Bowman, G. R., Beauchamp, K., Voelz, V. A. & Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **133**, 18413–18419 (2011).
20. Berezovska, G., Prada-Garcia, D. & Rao, F. Consensus for the Fip35 folding mechanism? *J. Chem. Phys.* **139**, 035102 (2013).
21. Wintjens, R. *et al.* ^{1}H NMR study on the binding of Pin1 Trp-Trp domain with phosphothreonine peptides. *J. Biol. Chem.* **276**, 25150–25156 (2001).
22. a Beccara, S., Škrbić, T., Covino, R. & Faccioli, P. Dominant folding pathways of a WW domain. *Proc. Natl. Acad. Sci. USA* **109**, 2330–2335 (2012).
23. McGibbon, R. T. & Pande, V. S. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theor. Comput.* **9**, 2900–2906 (2013).
24. Liu, F., Nakaema, M. & Gruebele, M. The transition state transit time of WW domain folding is controlled by energy landscape roughness. *J. Chem. Phys.* **131**, 195101 (2009).
25. Otting, G., Liepinsh, E. & Wüthrich, K. Disulfide bond isomerization in BPTI and BPTI(G36S): An NMR study of correlated mobility in proteins. *Biochemistry* **32**, 3571–3582 (1993).
26. Grey, M. J., Wang, C. & Palmer III, A. G. Disulfide bond isomerization in basic pancreatic trypsin inhibitor: Multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J. Am. Chem. Soc.* **125**, 14324–14335 (2003).
27. Long, D. & Brüschweiler, R. Atomistic kinetic model for population shift and allostery in biomolecules. *J. Am. Chem. Soc.* **133**, 18999–19005 (2011).
28. Xue, Y., Ward, J. M., Yuwen, T., Podkorytov, I. S. & Skrynnikov, N. R. Microsecond time-scale conformational exchange in proteins: Using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. *J. Am. Chem. Soc.* **134**, 2555–2562 (2012).
29. Noé, F., Wu, H., Prinz, J.-H. & Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **139**, 184114 (2013).
30. Vitalis, A. & Caflisch, A. Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theor. Comput.* **8**, 1108–1120 (2012).
31. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

Acknowledgments

The authors thank D.E. Shaw Research for sharing the trajectory data and their state annotation for BPTI (colour code in Fig. 2). We thank Dr. Francesco Rao for providing the state partitioning for FiP35 we use in Fig. 1. AV acknowledges financial support from the Holcim foundation. This work was supported in part by a grant from the Swiss National Science Foundation to AC.

Author contributions

N.B., A.V. and A.C. contributed to study design. N.B. analysed the data and created the figures and captions. A.V. wrote the manuscript, which was reviewed by all authors.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Blöchliger, N., Vitalis, A. & Caflisch, A. High-Resolution Visualisation of the States and Pathways Sampled in Molecular Dynamics Simulations. *Sci. Rep.* **4**, 6264; DOI:10.1038/srep06264 (2014).

Chapter 4

Peptide binding to a PDZ domain by electrostatic steering via non-native salt bridges

Blöchliger, N., Xu, M., and Caflisch, A. *Biophysical Journal*, 108(9): 2362–2370, 2015

Biophysical Journal Volume 108 May 2015 2362–2370

Article

Peptide Binding to a PDZ Domain by Electrostatic Steering via Nonnative Salt Bridges

Nicolas Blöchliger,[1] Min Xu,[1] and Amedeo Caflisch[1,*]
[1]Department of Biochemistry, University of Zurich, Zurich, Switzerland

ABSTRACT We have captured the binding of a peptide to a PDZ domain by unbiased molecular dynamics simulations. Analysis of the trajectories reveals on-pathway encounter complex formation, which is driven by electrostatic interactions between negatively charged carboxylate groups in the peptide and positively charged side chains surrounding the binding site. In contrast, the final stereospecific complex, which matches the crystal structure, features completely different interactions, namely the burial of the hydrophobic side chain of the peptide C-terminal residue and backbone hydrogen bonds. The simulations show that nonnative salt bridges stabilize kinetically the encounter complex during binding. Unbinding follows the inverse sequence of events with the same nonnative salt bridges in the encounter complex. Thus, in contrast to protein folding, which is driven by native interactions, the binding of charged peptides can be steered by nonnative interactions, which might be a general mechanism, e.g., in the recognition of histone tails by bromodomains.

INTRODUCTION

The fundamental process of protein-protein binding can be conceptualized as diffusional association followed by formation of the stereospecific complex (1–3). Long-range electrostatic forces can significantly accelerate and guide diffusional association, a phenomenon termed electrostatic steering (1,3–9). Association results in a relatively weak encounter complex, which is stabilized mainly by nonspecific interactions and whose binding interface is not yet fully desolvated (1,2,7,10,11). Crossing the transition state, potentially through multiple pathways (12), specific short-range hydrogen bonds and hydrophobic interactions form and the stereospecific complex is reached (1).

PDZ (PSD-95/Discs large/ZO-1) domains, which are found in scaffold proteins involved in signaling (13–15), have been used as model systems to study peptide binding (16,17). They share a common fold with six β-strands and two α-helices (Fig. 1 *A*) and mainly interact with target proteins by binding their C-termini (13,18), although binding to internal protein segments has been reported as well (19). In the stereospecific complex the side chain of a hydrophobic residue at the C-terminus of the target is buried and its carboxylate group interacts with the carboxylate-binding loop. In addition, backbone hydrogen bonds create an intermolecular β-sheet, and specificity is achieved by side-chain interactions (13,20,21).

In a previous molecular dynamics (MD) study on the third PDZ domain of the postsynaptic density protein 95, we focused the analysis on the binding site of the PDZ domain by comparing MD runs of the apo structure of the PDZ domain with MD runs started from the bound state (22). This comparison suggested that the peptide binds by conformational selection. No peptide dissociation event was observed because the length of each of the four trajectories started from the bound state was <0.2 μs. Several other MD simulations of PDZ domains have been performed during the last few years (23–27). However, we are not aware of any MD simulations of the binding of peptides to PDZ domains. Although unbiased MD simulations of sub-μs length have been used already to study the (reversible) binding of small and mainly rigid molecules to proteins (28–32), it is much more challenging to simulate the binding of flexible (oligo)peptides to proteins because the larger conformational space requires significantly longer trajectories (9,12,33).

Here, we report on unbiased, multiple MD simulations of 2.1–3.6 μs each, which were carried out to characterize the binding and unbinding of the C-terminal hexapeptide segment Acetyl-EQVSAV of the Ras-associating guanine nucleotide exchange factor 2 (RA-GEF2, also known as PDZ-GEF2 or RapGEF6) (34) to the second PDZ domain of protein tyrosine phosphatase 1E (PTP1E, also known as PTPL1, FAP-1, or PTP-Bas) (35,36). This study focuses on the intermolecular interactions during the (un)binding process and was motivated by the following questions. Is it possible to capture the spontaneous binding of a flexible hexapeptide to the PDZ domain by MD simulations on a conventional compute cluster? What is the role of the electrostatic interactions in the initial association and final binding? Does the binding proceed through native interactions, i.e., via the intermolecular contacts of the stereospecific complex as observed in the crystal structure? Is unbinding the reverse of binding?

Submitted February 20, 2015, and accepted for publication March 17, 2015.
*Correspondence: caflisch@bioc.uzh.ch
Editor: Rohit Pappu.

0006-3495/15/05/2362/9 $2.00
http://dx.doi.org/10.1016/j.bpj.2015.03.038

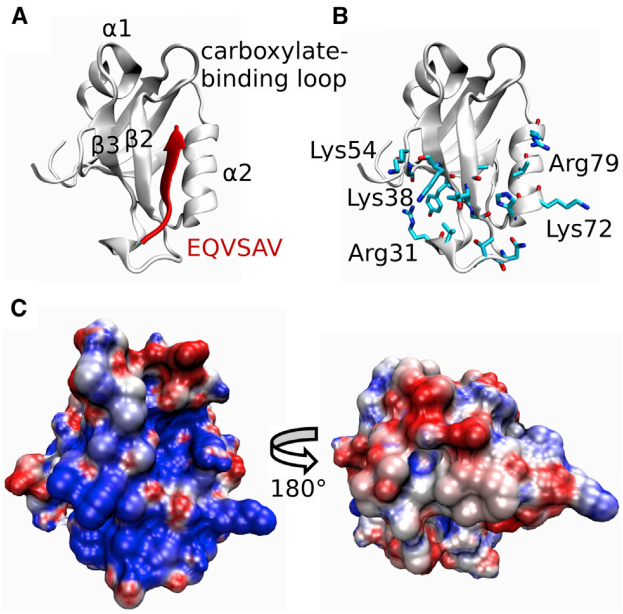


FIGURE 1 Initial association of the peptide and the PDZ domain by electrostatic steering. (*A*) Crystal structure of the stereospecific complex of PTP1E PDZ2 and the C-terminal RA-GEF2 peptide (PDB code 3LNY). The PDZ domain is shown in white with some secondary structure elements labeled. The peptide and its sequence are in red. (*B*) Surface of initial association. PDZ residues having an average contact frequency with the peptide residues of at least 0.1 during association are shown in a stick-like representation (see Materials and Methods). The structure and the orientation are the same as in (*A*). The contact frequency values are shown in Fig. S1. (*C*) Electrostatic surface potential. The color scale ranges from −5 kT/e (*red*) to 5 kT/e (*blue*). The orientation is the same as in (*A*) for the left panel. All illustrations were rendered with VMD (85). To see this figure in color, go online.

We observed multiple events of spontaneous binding in 10 MD runs of ~2.3 μs each started from fully unbound, and several rebinding events in 10 MD runs of ~3.4 μs each started from the crystal structure of the complex. Fast initial association is always driven and stabilized by long-range electrostatic interactions between negatively charged carboxylate groups in the peptide and positively charged side chains in the vicinity of the binding site. These salt bridges are not present in the final stereospecific complex.

MATERIALS AND METHODS

MD simulations

We carried out 10 independent simulations starting with the peptide placed randomly in the simulation box (called binding runs in the following) and 10 simulations started from the bound state (called unbinding runs). The total simulation length amounted to 57 μs.

The coordinates of PTP1E PDZ2 in complex with the C-terminal RA-GEF2 peptide were downloaded from the protein database (Protein Data Bank (PDB) code 3LNY, URL www.rcsb.org) (37). The sequence of the C-terminal RE-GEF2 peptide used here is EQVSAV, and its N-terminus was capped with acetyl. To reproduce neutral pH conditions, the side chains of aspartates and glutamates were negatively charged, those of lysines and arginines were positively charged, and the histidine side chains were neutral. The structure was solvated in a cubic water box. For the binding runs the peptide was placed randomly in the box with a resulting mean distance to the PDZ domain of 12 Å. The size of the box was 73 Å for the binding runs and 63 Å for the unbinding runs. The simulation system contained sodium and chloride ions to approximate an ionic strength of 150 mM and to compensate for the total charge of the two molecules. The simulations were carried out with GROMACS 4.5.5 (38) using the CHARMM27 force field (39,40) and the TIP3P water model (41). Periodic boundary conditions were applied, and electrostatic interactions were evaluated using the particle-mesh Ewald summation method (42). The van der Waals interactions were truncated at a cutoff of 10 Å. The temperature of 310 K was kept constant by an external bath with velocity rescaling (43), and the pressure was kept close to 1 atm by the Berendsen barostat (44). The LINCS algorithm was used to fix the covalent bonds involving hydrogen atoms (45). The integration time step was 2 fs, and snapshots were saved every 10 ps. Each MD run was carried out on 16 cores (i.e., four Xeon5560 CPUs) of the Schrödinger supercomputer at the University of Zurich, which required ~1 week per μs.

SAPPHIRE plot

Recently, we have developed an algorithm for the analysis of long MD trajectories (46,47). The resulting SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot is a comprehensive visualization of the thermodynamics and kinetics of the simulated system. A function measuring distance between snapshots is needed to generate SAPPHIRE plots and can be freely chosen by the user. We chose the Euclidean distance function on 29 distances between atoms of the peptide and the binding site of the PDZ domain for the present application. Table S1 in the Supporting Material contains the full list of atom pairs used.

We briefly describe the method here and refer the reader to the original publications for more details (46,47). Starting from an arbitrary snapshot, all the snapshots are sequentially ordered in a stepwise fashion. In each step, the snapshot closest to any snapshot prior in the sequence becomes the next entry. The complete sequence of snapshots is called progress index. Assuming high snapshot density within free energy basins, snapshots belonging to the same basin are grouped together and distinct states do not overlap (46). A stochastic algorithm to generate an approximate progress index has been developed. This algorithm is scalable to large data sets and was used here. It is important to note that the progress index is not a reaction coordinate. It is rather a sorting of all MD snapshots to identify basins without any a priori clustering.

We employ three types of annotation functions to highlight and interpret the states along the progress index and the pathways connecting them (Fig. 2). First, we use a kinetic annotation function to localize the individual states on the progress index. Specifically, for every snapshot i along the progress index, we plot the average of the mean first-passage times between A_i and S_i, denoted τ_{MFP}, where A_i is the set of snapshots added to the progress index before i and S_i is the set of those added after i. The value of this annotation function is low within a state and high in transition regions, and barriers are highlighted reliably (although they cannot be interpreted quantitatively) (46). Second, we plot the actual sampling time of the individual snapshots to illustrate when and in which sequence the different states were sampled. This information appears as red dots in Fig. 2 and corresponds to the trace of the temporal evolution of the system, i.e., the detailed sequence of events for each MD run. Third, we characterize the states themselves by a structural annotation. In this case we have used the distance between the peptide and the PDZ domain, the solvent accessible surface area of the peptide, the root mean-square deviation (RMSD) of the peptide with respect to a reference structure after alignment on the PDZ domain, as well as several interatomic distances.

Trajectories from the individual simulation runs were concatenated and subsampled at 20 ps to generate the SAPHHIRE plot. For the unbinding runs, the size of the simulation box was adjusted to match the binding runs after the system has been centered on the PDZ atoms. The stochastic

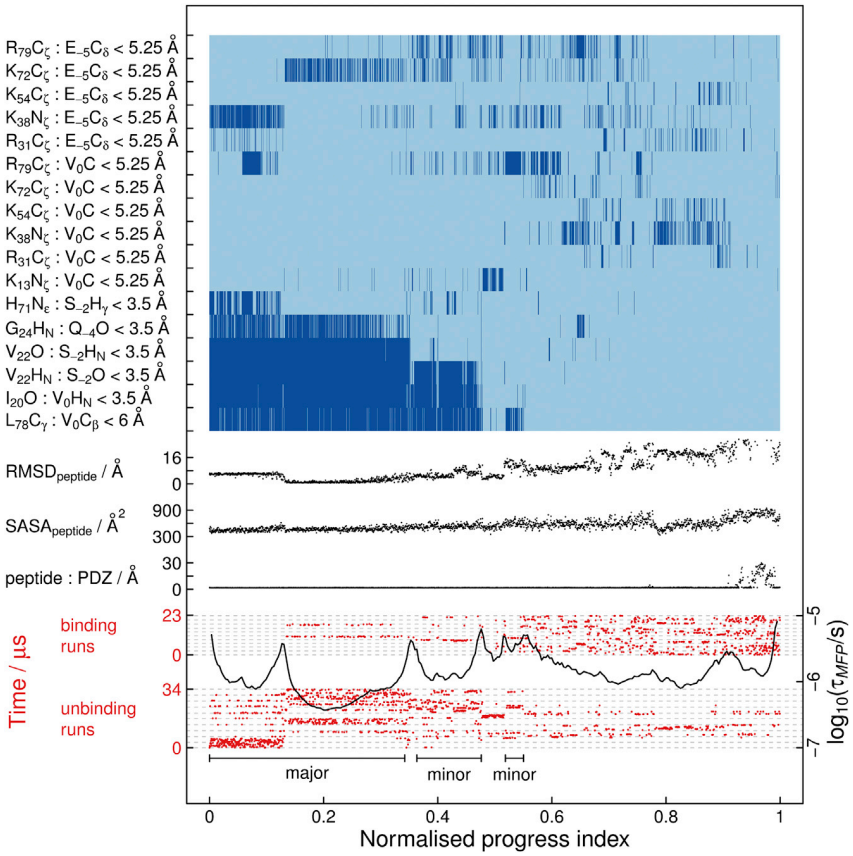


FIGURE 2 SAPPHIRE plot illustrating the sampled ensemble. The progress index (x axis) represents a reordering of the trajectory snapshots that groups similar snapshots next to each other (see Materials and Methods). The progress index is annotated with kinetic information (τ_{MFP}, a function whose value is low within states and high in transition regions, *black profile* in the *bottom*), sampling time (*red dots*), and structural information (*middle* and *top*). The annotation in the top part of the panel uses binning, with dark blue meaning that the distance given on the left side (reporting on burial of the Val0 side chain, intermolecular backbone hydrogen bonds, and salt bridges) is below the indicated threshold. $RMSD_{peptide}$ was computed on the C_α atoms of the peptide with respect to a representative structure of the major binding mode after alignment on the PDZ domain. $SASA_{peptide}$ is the solvent accessible surface area of the peptide. Peptide: PDZ denotes the minimal distance between the peptide and the PDZ domain. Gray dashed lines indicate the boundaries between individual simulation runs. The major and minor binding modes of the stereospecific complex are labeled (*black horizontal segments* in the *bottom*). To see this figure in color, go online.

algorithm mentioned previously is scalable because of the preorganization of the data via tree-based, hierarchical clustering (48). The lower and upper threshold radius and the tree height for the clustering were set to 0.6 Å, 10 Å, and 12. The first snapshot on the progress index is the starting structure of the first run, i.e., the crystal structure of the bound complex (49). The number of guesses to find nearest neighbors (46) was set to 10^4. The method is implemented in the CAMPARI simulation and analysis package (http://campari.sourceforge.net).

Contact frequencies

First, the MD trajectory segments were classified as stereospecific complex or other, where other includes fully unbound and encounter complex. For this binary classification the kinetic annotation of the SAPPHIRE plot was used (*black profile* in Fig. 2), as well as the RMSD of the peptide with respect to the major binding mode, and various distances between the peptide and the binding pocket (Figs. S2–S21). This classification is illustrated in the top of Figs. S2–S21. The 10 binding runs were then used to compute contact frequencies by employing only the segments annotated as other (i.e., fully unbound and encounter complex). A contact is considered to be formed between a residue of the peptide and a residue of the PDZ domain if two atoms are within 5 Å. The acetyl at the N-terminus of the peptide was considered as an independent residue, and CAMPARI (http://campari.sourceforge.net) was used for this analysis.

Electrostatic surface potential

The electrostatic potential on the surface of the PDZ domain was calculated with PDB2PQR (50,51) and APBS (52) using the conformation of the PDZ domain in the stereospecific complex (PDB entry 3LNY).

Binding time and k_{on}

Mean binding times were separately estimated for the binding runs and the rebinding events observed in the unbinding runs as $\tau = t_{unbound}/n$, where $t_{unbound}$ is the total time the peptide is not bound as in the stereospecific complex (defined previously and annotated in Figs. S2–S21) and n is the number of binding events. For the binding runs $n = 5$ and $\tau = 3.8$ μs. We observed $n = 3$ rebinding events in the unbinding runs, resulting in $\tau = 2.4$ μs. The rate constant k_{on} was estimated to be $1/\tau$ [peptide], where [peptide] = 4.3 mM and 6.7 mM for the binding and unbinding runs, respectively. The resulting values for k_{on} are 61 $\mu M^{-1}s^{-1}$ and 63 $\mu M^{-1}s^{-1}$ for the binding and unbinding runs, respectively.

Free energy profile

Cut-based (53,54) and conventional, histogram-based free energy profiles were computed using Fep1d (55).

RESULTS AND DISCUSSION

We performed 10 simulations starting from the peptide placed randomly in the simulation box. These 10 simulations (called binding runs in the following) were completely agnostic of the binding site, and no biasing force or restraint was used. In addition 10 independent runs were started from the bound state using the crystal structure of the complex (PDB code 3LNY (49)) as starting conformation and different random seeds for the

initial assignment of the velocities. Total simulation time amounted to 57 μs.

Association via electrostatic steering and nonnative salt bridges

We observed fast association of the peptide and the PDZ domain in all of the binding runs. The intermolecular distance dropped below 2.5 Å within 5 ns on average. The residues most involved in complex formation are Val22, Thr23, Gly24, His71, Val75, and Arg79 located along the binding site, Asn27, Thr28, Val30, Arg31, Tyr36, and Lys38 in the β2-β3 loop and on strand β3, as well as Lys54, Gly55, and Lys72 (Figs. 1 *B* and S1 and Materials and Methods). On the other hand, contact frequencies are low for the carboxylate binding loop, for helix α1, which contains two negatively charged residues, and the β-sheet formed by the β1, β6, β4, and β5 strands, which is located on the other side of the domain with respect to the binding site. The electrostatic potential on this surface of initial association is positively charged (Fig. 1 *C*), and diffusion of the peptide, which bears two negative charges, to the vicinity of the binding site is thus mainly driven by electrostatic steering.

Various salt bridges are formed in the encounter complex, which features multiple relative orientations of the peptide and PDZ domain. In the fifth binding run, for example, the carboxylate group of Val0 (peptide residues are numbered from −5 to 0) forms salt bridges with Arg79, Lys13, and Lys72 before committing to the final binding pose (Fig. 3 and Movie S1). The detailed sequence of events and the roles played by the individual charged residues are different in the other simulation runs (Figs. S2–S21), illustrating the heterogeneity of the encounter complex and the lack of specific interactions (10,11,56–58). However, in all of the binding runs salt bridges are dynamically formed in the encounter complex.

The solvent accessible surface area of the peptide is larger in the encounter complex than in the stereospecific complex (Figs. 3 and S2–S21). This indicates that the binding interface of the encounter complex is not fully desolvated and that specific intermolecular hydrogen bonds or hydrophobic interactions are of lesser importance (10,56,57).

Final binding in antiparallel β-sheet arrangement

We have recently developed a method for the visualization of long MD trajectories (46,47). The main output of the method is the SAPPHIRE plot, which offers an intuitive illustration of the states and sequence of events encountered during the simulation (see Materials and Methods). Previously, we used SAPPHIRE plots to analyze protein folding and conformational changes in the native state of a protein (47) as well as multiple conformations of a loop of the prion protein (59). Here, we apply the method to a binding process.

The SAPPHIRE plot of the combined binding and unbinding runs (Fig. 2) shows that the major binding mode is stabilized by the canonical burial of the Val0 side chain (formed contact with side chain of Leu78 in the binding pocket). In addition, the Val0 carboxylate group interacts with the carboxylate-binding loop (Fig. 3 *A*) and intermolecular backbone hydrogen bonds are formed between the carbonyl

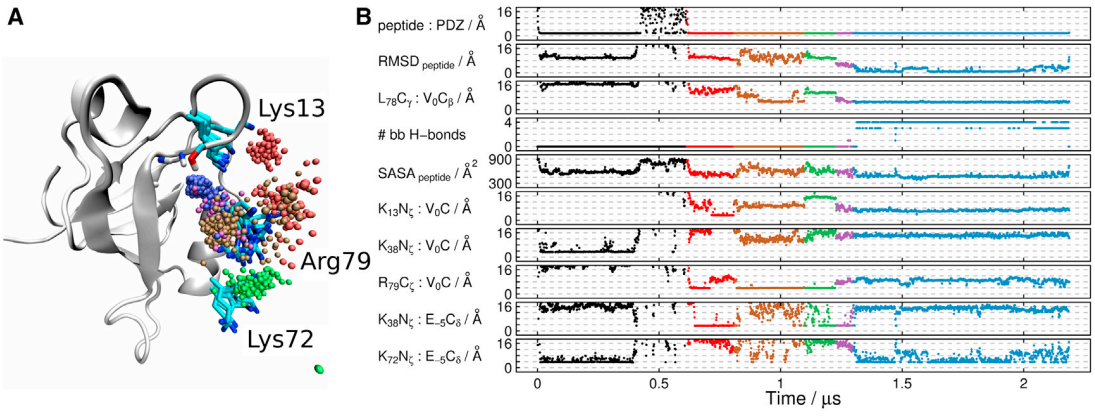


FIGURE 3 Salt bridges stabilize the encounter complex. (*A*) Nonspecific salt bridges in the encounter complex during binding in the fifth binding run (Movie S1). A representative conformation of the PDZ domain in the major binding mode is shown along with the backbone amide groups of the carboxylate-binding loop and the side chain of Ser17. A sphere is drawn every 1 ns at the positon of the carbon atom of the Val0 carboxylate group during binding and colored according to time, as indicated in (*B*). The side chains of selected basic residues involved in salt bridges with the carboxylate group of Val0 are drawn every 150 ns. The illustration was rendered with VMD (85). (*B*) Analysis of the fifth binding run. The minimal distance between the peptide and the PDZ domain (peptide: PDZ), the RMSD of the peptide C_α atoms after alignment on the PDZ domain, the number of backbone hydrogen bonds formed (i.e., distances between the carbonyl oxygen of Ile20 and the NH group of Val0, between both polar groups of Val22 and Ser-2, and between the NH group of Gly24 and the carbonyl oxygen of Gln-4 below 3.5 Å), the solvent accessible surface area of the peptide ($SASA_{peptide}$), and distances between selected atom pairs are plotted as median values in a window of 1 ns. Corresponding plots for the other simulation runs are given in Figs. S2–S21. To see this figure in color, go online.

oxygen of Ile20 and the NH group of Val0, between both polar groups of Val22 and Ser-2, and between the NH group of Gly24 and the carbonyl oxygen of Gln-4 (Fig. 2). Of importance, the most populated binding mode is essentially identical to the crystal structure. The barrier at a value of the normalized progress index of ~0.15 is due to reorientation of the Glu-5 side chain, which can either point toward the solvent or form a salt bridge with Lys72. Regarding the crystal structure, note that the atoms of the Glu-5 residue had very high B-factors and the side chain did not show any electron density (49). Furthermore, the peptide used by Zhang et al. is slightly longer than the one we simulated and has an additional charged residue (Glu-7), which is likely to affect, at least in part, the orientation of the N-terminal segment of the peptide in the bound conformation.

A minor binding mode is located between normalized progress index values of ~0.36 and ~0.48 (Fig. 2). In this binding mode the C-terminal part of the peptide is bound as in the crystal structure, whereas the N-terminal segment protrudes into the solvent. Only the two backbone hydrogen bonds toward the C-terminus of the peptide (between the carbonyl oxygen of Ile20 and the NH group of Val0 and between the NH group of Val22 and the carbonyl oxygen of Ser-2) are formed, in agreement with recent experimental results obtained by amide-to-ester mutations (57). Another alternative binding mode, which is short-lived and was repeatedly sampled, is located between normalized progress index values of ~0.52 and ~0.55. This binding mode features burial of the Val0 side chain (as in the crystal structure), whereas the Val0 carboxylate group forms a salt bridge with Arg79 instead of interacting with the carboxylate-binding loop and no intermolecular backbone hydrogen bonds are present. Finally, snapshots representing the encounter complex are found between normalized progress index values of ~0.55 and ~0.9. Fully unbound conformations accumulate at the end of the progress index.

The stereospecific complex (major or minor binding modes) was reached in five out of the 10 binding runs (Figs. 2, 3, and S2–S11). Additionally, the peptide rebound in three of the six unbinding runs in which full dissociation was observed (Figs. S12–S21 and Movie S2). Our estimate for k_{on} based on these eight binding events is ~60 $\mu M^{-1}s^{-1}$ (see Materials and Methods). We note that the TIP3P water model used here shows a self-diffusion constant higher by a factor of 2–3 than the experimentally measured value (60), which might influence k_{on}. Experimental values for k_{on} collected at lower temperatures and similar or higher ionic strength range from 2.9 to 36 $\mu M^{-1}s^{-1}$ for the same PDZ domain or its mouse ortholog PTP-BL PDZ2 and the peptide ENEQVSAV or *dansyl*-EQVSAV (49,57,61–63).

The dissociation of the encounter complex is frequent on the timescale of binding in our simulations as the average lifetime of the encounter complex is ~200 ns (see distance between peptide and PDZ domain in Figs. S2–S21). The encounter complex is thus located before the rate-limiting step (2,56,57). This observation is validated by the free energy profile along the distance between the Val0 side chain and the hydrophobic pocket of the PDZ domain (Fig. 4), which confirms that the main barrier accounts for the burial of the Val0 side chain. Comparing Fig. 3 with the corresponding figures for the other simulation runs (Figs. S2–S21) shows that the stereospecific complex can be reached from the encounter complex via various pathways. The burial of the Val0 side chain takes place before the formation of the backbone hydrogen bonds or almost simultaneously (e.g., in the binding run 8, Fig. S19). Thus, the sequence of events for binding starts with the formation of nonnative salt bridges in the encounter complex (which does not always lead to full binding) followed by burial of the Val0 side chain, and formation of the backbone hydrogen bonds between residues Val0/Ser-2 and the PDZ $\beta 2$ strand in an antiparallel β-sheet arrangement.

Inverse sequence of events during unbinding

It is interesting to analyze the unbinding process and compare with binding. Peptide dissociation starts by the rupture of the backbone hydrogen bonds, which takes place before the Val0 side chain exits from the hydrophobic pocket of the PDZ domain. Thus, the initial events of unbinding are the reverse of the final events of binding. Furthermore, the peptide does not immediately diffuse away from the PDZ domain after the native interactions of the stereospecific complex break apart. Instead, the peptide remains in contact with the PDZ domain for several hundred nanoseconds (Movie S2 and Figs. S4, S5, S7–S9, and S21). Quantitatively, the residence time in the encounter complex is 650 ± 900 ns during unbinding and 200 ± 300 ns during binding. Of importance, the same nonnative salt bridges provide kinetic stabilization to the encounter complex during

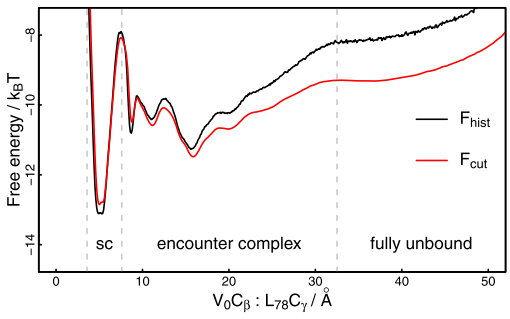


FIGURE 4 Free energy profile along a geometric order parameter. Histogram-based (*black*) and cut-based (*red*) (53,54) free energy profiles are shown as a function of the distance between the C_β atom of Val0 and the C_γ atom of Leu78, which reports on burial of the Val0 side chain. Barriers separating the stereospecific complex (sc), the encounter complex, and fully unbound conformations are indicated by gray, dashed lines. Note that this simple projection introduces overlap and hides crucial information, which, in contrast, is fully resolved by the SAPPHIRE plot (Fig. 2), e.g., the presence of minor binding modes. To see this figure in color, go online.

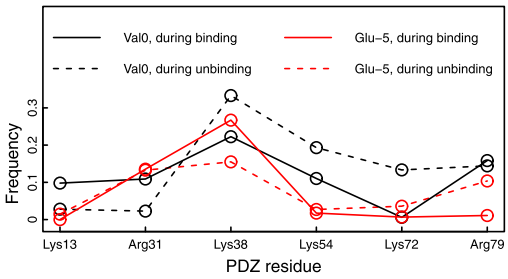


FIGURE 5 Salt bridges in the encounter complex during binding and unbinding. The trajectory segments that correspond to the encounter complex during binding and unbinding were extracted based on Figs. S2–S21, and salt bridges were considered to be formed if the N_ζ atom (for the PDZ lysines) or the C_δ atom (for the PDZ arginines) was within 6 Å of the carboxylate carbon of Val0 or Glu-5, respectively. To see this figure in color, go online.

both peptide association and dissociation (Fig. 5). Thus, the sequence of events for full dissociation is the reverse of binding.

CONCLUSIONS

We have used unbiased MD simulations to analyze the binding of the C-terminal hexapeptide segment of a natural ligand to the second PDZ domain of PTP1E. The general view of the binding process is schematically depicted in Fig. 6 and a representative binding event is shown in Movie S1. Initial association is driven by the long-range electrostatic interactions between the peptide and the PDZ domain (Fig. 1 *C*). In the resulting encounter complex the peptide is weakly bound in the vicinity of the binding site (Figs. 1 *B* and 3). The complex is maintained by nonspecific electrostatic interactions, which allows the peptide to sample multiple orientations (Fig. 3). After the rate-limiting step the side chain of Val0 is buried in a hydrophobic pocket (Figs. 2 and 3). At this point, up to four backbone hydrogen bonds between the peptide and β2 can form depending on whether the major binding mode is reached directly or via distinct minor binding modes (Fig. 2). The comparison of the sequence of events for binding and unbinding shows that the two processes are one the inverse of the other.

To further investigate the influence of the encounter complex on the rate constant for binding, we suggest to measure experimentally the salt dependence of the binding rate, e.g., by the Förster resonance energy transfer technique. These measurements have already been reported for PTP-BL PDZ2 and a dansylated peptide without any charged side chains (64). Whereas k_{off} was independent of the ionic strength, k_{on} decreased with increasing ionic strength, which was attributed to the negative charge of the C-terminal carboxylate group. A stronger influence on k_{on} is predicted, on the basis of our MD simulation results, for a similar peptide with one or two negatively charged side chains. On the other hand, electrostatic steering has been ruled out for binding of a peptide with no net charge (*dansyl*-KQTSV) to PDZ3 of postsynaptic density protein 95 (which has glutamic acids at the positions of Arg31 and Lys72) (65).

The mechanism of initial association guided by nonspecific electrostatic steering is likely to be valid for other (small, single-domain) peptide-binding proteins (4,7,9,66,67). As an example, the binding of histone tails to bromodomains is most probably driven by the negative electrostatic potential on the surface surrounding the acetylated lysine binding site, whereas the final stereospecific complex is stabilized by the hydrogen bond between the acetyl carbonyl and the side chain of the evolutionary conserved Asn (68). Other examples include the binding of phosphorylated peptides to SH2 domains (12) as well as intrinsically disordered proteins (69), which tend to contain more charged residues than globular proteins (70). Regarding the coupled binding and folding of intrinsically disordered proteins (71), experimental and theoretical (72,73) studies have highlighted nonnative salt bridges in the encounter complex (74), enhanced on-rates due to electrostatic interactions (75–77), nonnative steering (78), and late formation of native contacts (79).

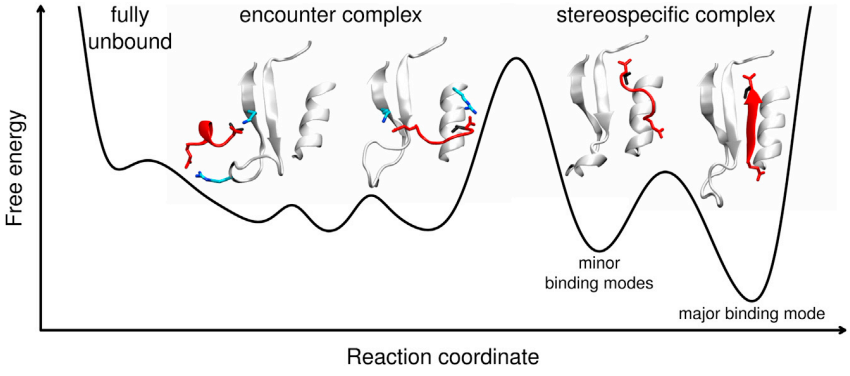


FIGURE 6 Schematic free energy profile of the binding process. After association accelerated by electrostatic steering a weak encounter complex is formed, which is stabilized by nonspecific intermolecular salt bridges. In contrast, the stereospecific complex features burial of the Val0 side chain and multiple binding modes differing among each other only in the orientation of the N-terminal part of the peptide. In this qualitative illustration, relative barrier heights roughly reflect the kinetics observed in the MD simulations, i.e., fast formation, reconfiguration and dissociation of the encounter complex, interconversions among major and minor binding modes on an intermediate timescale, and slow transitions between encounter complex and stereospecific complex. (*Insets*) The ribbon illustrations focus on the binding site, i.e., only the following structural elements are shown for clarity: Carboxylate-binding loop, β2-β3 strands and loop, and helix α2 of the PDZ domain (*gray*), backbone of the peptide, C-terminal carboxylate group, and the Glu-5 side (*red*), Val0 side chain (*black*), and the side chains of basic residues of the PDZ domain involved in salt bridges with the peptide (*cyan*). The ribbon illustrations were prepared with VMD (85). To see this figure in color, go online.

Finally, it is interesting to compare protein folding with peptide binding as they differ in the number of molecules involved but they are both governed by noncovalent interactions. Protein folding is driven by progressive formation of native interactions, which are in general more favorable than nonnative contacts (80–84). In contrast, our simulation results provide evidence that the binding of a charged peptide to a protein surface with opposite charge can be steered by long-range polar interactions that are not present in the final bound state.

SUPPORTING MATERIAL

Twenty-one figures, one table, and two movies are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)00299-4.

AUTHOR CONTRIBUTIONS

All authors contributed to the study design. M.X. performed the MD simulations and prepared the Supplementary Movies, N.B. and A.C. analyzed the data and wrote the article.

ACKNOWLEDGMENTS

This work was supported by a grant from the Swiss National Science Foundation to A.C.

REFERENCES

1. Schreiber, G., G. Haran, and H. X. Zhou. 2009. Fundamental aspects of protein-protein association kinetics. *Chem. Rev.* 109:839–860.
2. Schreiber, G. 2002. Kinetic studies of protein-protein interactions. *Curr. Opin. Struct. Biol.* 12:41–47.
3. Berg, O. G., and P. H. von Hippel. 1985. Diffusion-controlled macromolecular interactions. *Annu. Rev. Biophys. Biophys. Chem.* 14:131–160.
4. Schreiber, G., and A. R. Fersht. 1996. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* 3:427–431.
5. Selzer, T., S. Albeck, and G. Schreiber. 2000. Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.* 7:537–541.
6. Hemsath, L., R. Dvorsky, …, M. R. Ahmadian. 2005. An electrostatic steering mechanism of Cdc42 recognition by Wiskott-Aldrich syndrome proteins. *Mol. Cell.* 20:313–324.
7. Northrup, S. H., J. O. Boles, and J. C. L. Reynolds. 1988. Brownian dynamics of cytochrome *c* and cytochrome *c* peroxidase association. *Science.* 241:67–70.
8. Gabdoulline, R. R., and R. C. Wade. 1997. Simulation of the diffusional association of barnase and barstar. *Biophys. J.* 72:1917–1929.
9. Ahmad, M., W. Gu, and V. Helms. 2008. Mechanism of fast peptide recognition by SH3 domains. *Angew. Chem. Int.Ed. Engl.* 47:7626–7630.
10. Tang, C., J. Iwahara, and G. M. Clore. 2006. Visualization of transient encounter complexes in protein-protein association. *Nature.* 444:383–386.
11. Suh, J. Y., C. Tang, and G. M. Clore. 2007. Role of electrostatic interactions in transient encounter complexes in protein-protein association investigated by paramagnetic relaxation enhancement. *J. Am. Chem. Soc.* 129:12954–12955.
12. Giorgino, T., I. Buch, and G. De Fabritiis. 2012. Visualizing the induced binding of SH2-phosphopeptide. *J. Chem. Theory Comput.* 8:1171–1175.
13. Sheng, M., and C. Sala. 2001. PDZ domains and the organization of supramolecular complexes. *Annu. Rev. Neurosci.* 24:1–29.
14. Harris, B. Z., and W. A. Lim. 2001. Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* 114:3219–3231.
15. Feng, W., and M. Zhang. 2009. Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. *Nat. Rev. Neurosci.* 10:87–99.
16. Jemth, P., and S. Gianni. 2007. PDZ domains: folding and binding. *Biochemistry.* 46:8701–8708.
17. Chi, C. N., A. Bach, …, P. Jemth. 2012. Ligand binding by PDZ domains. *Biofactors.* 38:338–348.
18. Doyle, D. A., A. Lee, …, R. MacKinnon. 1996. Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell.* 85:1067–1076.
19. Hillier, B. J., K. S. Christopherson, …, W. A. Lim. 1999. Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science.* 284:812–815.
20. Songyang, Z., A. S. Fanning, …, L. C. Cantley. 1997. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science.* 275:73–77.
21. Stiffler, M. A., J. R. Chen, …, G. MacBeath. 2007. PDZ domain binding selectivity is optimized across the mouse proteome. *Science.* 317:364–369.
22. Steiner, S., and A. Caflisch. 2012. Peptide binding to the PDZ3 domain by conformational selection. *Proteins: Struct., Funct. Bioinf.* 80:2562–2572.
23. Kong, Y., and M. Karplus. 2009. Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins: Struct., Funct. Bioinf.* 74:145–154.
24. Mostarda, S., D. Gfeller, and F. Rao. 2012. Beyond the binding site: the role of the β_2-β_3 loop and extra-domain structures in PDZ domains. *PLOS Comput. Biol.* 8:e1002429.
25. Buchli, B., S. A. Waldauer, …, P. Hamm. 2013. Kinetic response of a photoperturbed allosteric protein. *Proc. Natl. Acad. Sci. USA.* 110:11725–11730.
26. Buchenberg, S., V. Knecht, …, G. Stock. 2014. Long-range conformational transition of a photoswitchable allosteric protein: molecular dynamics simulation study. *J. Phys. Chem. B.* 118:13468–13476.
27. Dhulesia, A., J. Gsponer, and M. Vendruscolo. 2008. Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein. *J. Am. Chem. Soc.* 130:8931–8939.
28. Shan, Y., E. T. Kim, …, D. E. Shaw. 2011. How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* 133:9181–9183.
29. Dror, R. O., A. C. Pan, …, D. E. Shaw. 2011. Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA.* 108:13118–13123.
30. Huang, D., and A. Caflisch. 2011. Small molecule binding to proteins: affinity and binding/unbinding dynamics from atomistic simulations. *ChemMedChem.* 6:1578–1580.
31. Gohlke, H., U. Hergert, …, L. Schmitt. 2013. Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *J. Chem. Inf. Model.* 53:2493–2498.
32. Buch, I., T. Giorgino, and G. De Fabritiis. 2011. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA.* 108:10184–10189.
33. Magno, A., S. Steiner, and A. Caflisch. 2013. Mechanism and kinetics of acetyl-lysine binding to bromodomains. *J. Chem. Theory Comput.* 9:4225–4232.
34. Gao, X., T. Satoh, …, T. Kataoka. 2001. Identification and characterization of RA-GEF-2, a Rap guanine nucleotide exchange factor that serves as a downstream target of M-Ras. *J. Biol. Chem.* 276:42219–42225.

35. Erdmann, K. S. 2003. The protein tyrosine phosphatase PTP-Basophil/Basophil-like. Interacting proteins and molecular functions. *Eur. J. Biochem.* 270:4789–4798.

36. Abaan, O. D., and J. A. Toretsky. 2008. PTPL1: a large phosphatase with a split personality. *Cancer Metastasis Rev.* 27:205–214.

37. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

38. Van Der Spoel, D., E. Lindahl, …, H. J. C. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.

39. MacKerell, Jr., A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.

40. Mackerell, Jr., A. D., M. Feig, and C. L. Brooks, 3rd. 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25:1400–1415.

41. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

42. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

43. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.

44. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

45. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.

46. Blöchliger, N., A. Vitalis, and A. Caflisch. 2013. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.* 184:2446–2453.

47. Blöchliger, N., A. Vitalis, and A. Caflisch. 2014. High-resolution visualization of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.* 4:6264.

48. Vitalis, A., and A. Caflisch. 2012. Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theory Comput.* 8:1108–1120.

49. Zhang, J., P. J. Sapienza, …, A. L. Lee. 2010. Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of protein tyrosine phosphatase 1E. *Biochemistry.* 49:9280–9291.

50. Dolinsky, T. J., J. E. Nielsen, …, N. A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32:W665–W667.

51. Dolinsky, T. J., P. Czodrowski, …, N. A. Baker. 2007. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* 35:W522–W525.

52. Baker, N. A., D. Sept, …, J. A. McCammon. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA.* 98:10037–10041.

53. Krivov, S. V., and M. Karplus. 2006. One-dimensional free-energy profiles of complex systems: progress variables that preserve the barriers. *J. Phys. Chem. B.* 110:12689–12698.

54. Krivov, S. V., and M. Karplus. 2008. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. USA.* 105:13841–13846.

55. Banushkina, P. V., and S. V. Krivov. 2015. Fep1d: a script for the analysis of reaction coordinates. *J. Comput. Chem.* Published online February 25, 2015. http://dx.doi.org/10.1002/jcc.23868.

56. Haq, S. R., C. N. Chi, …, P. Jemth. 2012. Side-chain interactions form late and cooperatively in the binding reaction between disordered peptides and PDZ domains. *J. Am. Chem. Soc.* 134:599–605.

57. Eildal, J. N. N., G. Hultqvist, …, P. Jemth. 2013. Probing the role of backbone hydrogen bonds in protein-peptide interactions by amide-to-ester mutations. *J. Am. Chem. Soc.* 135:12998–13007.

58. Volkov, A. N., J. A. R. Worrall, …, M. Ubbink. 2006. Solution structure and dynamics of the complex between cytochrome *c* and cytochrome *c* peroxidase determined by paramagnetic NMR. *Proc. Natl. Acad. Sci. USA.* 103:18945–18950.

59. Huang, D., and A. Caflisch. 2015. Evolutionary conserved Tyr-169 stabilizes the $\beta2$-$\alpha2$ loop of the prion protein. *J. Am. Chem. Soc.* 137:2948–2957.

60. Mahoney, M. W., and W. L. Jorgensen. 2001. Diffusion constant of the TIP5P model of liquid water. *J. Chem. Phys.* 114:363–366.

61. Gianni, S., A. Engström, …, P. Jemth. 2005. The kinetics of PDZ domain-ligand interactions and implications for the binding mechanism. *J. Biol. Chem.* 280:34805–34812.

62. Gianni, S., T. Walma, …, G. W. Vuister. 2006. Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure.* 14:1801–1809.

63. Gianni, S., S. R. Haq, …, P. Jemth. 2011. Sequence-specific long range networks in PSD-95/discs large/ZO-1 (PDZ) domains tune their binding selectivity. *J. Biol. Chem.* 286:27167–27175.

64. Di Silvio, E., D. Bonetti, …, S. Gianni. 2014. The mechanism of binding of the second PDZ domain from the Protein Tyrosine Phosphatase-BL to the Adenomatous Polyposis Coli tumor suppressor. *Protein Eng. Des. Sel.* 27:249–253.

65. Chi, C. N., A. Engström, …, P. Jemth. 2006. Two conserved residues govern the salt and pH dependencies of the binding reaction of a PDZ domain. *J. Biol. Chem.* 281:36811–36818.

66. Honig, B., and A. Nicholls. 1995. Classical electrostatics in biology and chemistry. *Science.* 268:1144–1149.

67. Sheinerman, F. B., R. Norel, and B. Honig. 2000. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* 10:153–159.

68. Filippakopoulos, P., S. Picaud, …, S. Knapp. 2012. Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell.* 149:214–231.

69. Dyson, H. J., and P. E. Wright. 2005. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6:197–208.

70. Uversky, V. N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11:739–756.

71. Dyson, H. J., and P. E. Wright. 2002. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12:54–60.

72. Baker, C. M., and R. B. Best. 2014. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4:182–198.

73. Chen, T., J. Song, and H. S. Chan. 2014. Theoretical perspectives on nonnative interactions and intrinsic disorder in protein folding and binding. *Curr. Opin. Struct. Biol.* 30:32–42.

74. Zhang, W., D. Ganguly, and J. Chen. 2012. Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLOS Comput. Biol.* 8:e1002353.

75. Rogers, J. M., A. Steward, and J. Clarke. 2013. Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited'. *J. Am. Chem. Soc.* 135:1415–1422.

76. Ganguly, D., S. Otieno, …, J. Chen. 2012. Electrostatically accelerated coupled binding and folding of intrinsically disordered proteins. *J. Mol. Biol.* 422:674–684.

77. Ganguly, D., W. Zhang, and J. Chen. 2013. Electrostatically accelerated encounter and folding for facile recognition of intrinsically disordered proteins. *PLOS Comput. Biol.* 9:e1003363.

78. De Sancho, D., and R. B. Best. 2012. Modulation of an IDP binding mechanism and rates by helix propensity and non-native interactions: association of HIF1α with CBP. *Mol. Biosyst.* 8:256–267.

79. Dogan, J., X. Mu, …, P. Jemth. 2013. The transition state structure for coupled binding and folding of disordered protein domains. *Sci. Rep.* 3:2076.

80. Dobson, C. M., A. Šali, and M. Karplus. 1998. Protein folding: a perspective from theory and experiment. *Angew. Chem. Int.Ed. Engl.* 37:868–893.

81. Dill, K. A., and H. S. Chan. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10–19.

82. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.

83. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.

84. Best, R. B., G. Hummer, and W. A. Eaton. 2013. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. USA.* 110:17874–17879.

85. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28.

Supporting Material to
Peptide Binding to a PDZ Domain by Electrostatic Steering via Non-Native Salt Bridges

Nicolas Blöchliger, Min Xu, and Amedeo Caflisch
Department of Biochemistry
University of Zurich
Winterthurerstrasse 190, CH-8057 Zurich

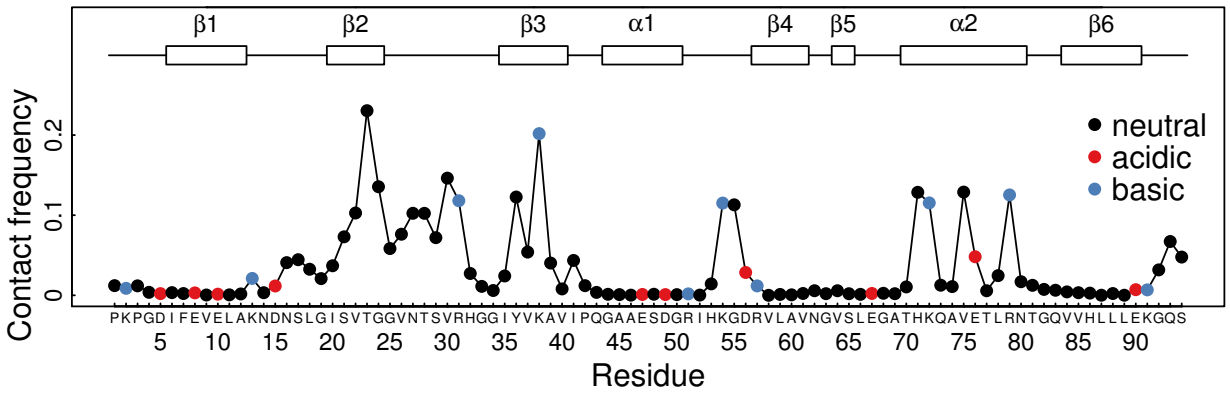


Fig. S1 Contact frequencies between protein residues and the peptide. For every residue, the average contact frequency over the peptide residues is plotted. Secondary structure elements of the protein are indicated on top. See Methods for details.

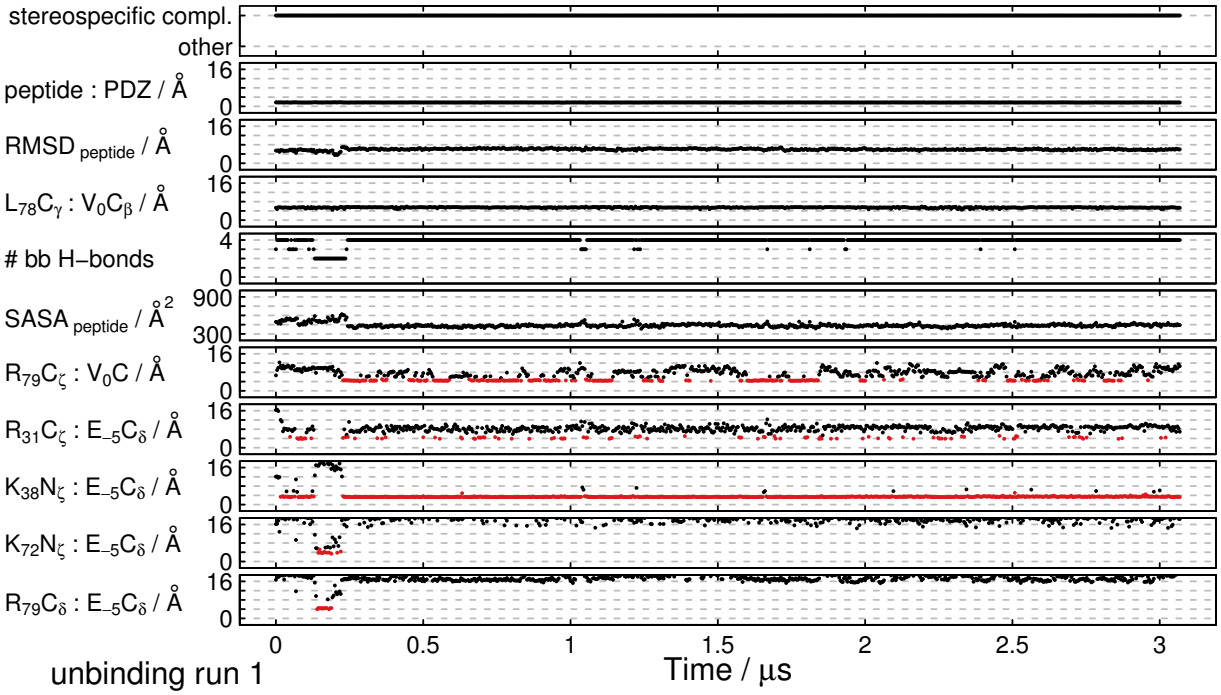


Fig. S2 Analysis of the first unbinding run. Distance between peptide and PDZ domain (peptide : PDZ), RMSD of the peptide C_α atoms after alignment on the PDZ domain, number of backbone hydrogen bonds formed (i.e. distances between the carbonyl oxygen of Ile20 and the NH group of Val0, between both polar groups of Val22 and Ser-2, and between the NH group of Gly24 and the carbonyl oxygen of Gln-4 below 3.5 Å), solvent accessible surface area of the peptide ($SASA_{peptide}$) and distances between selected atom pairs are shown. The distance between the Leu78 C_γ atom and the Val0 C_β atom reports on burial of the Val0 side chain in the binding pocket. Distances indicating salt bridge formation are coloured red if their value is below 5.25 Å. All quantities are shown as median values in a window of 1 ns. The top panel indicates the snapshots we classified as representing the stereospecific complex. Corresponding figures are given in Supplementary Figs. 3–11 for the other unbinding runs and in Supplementary Figs. 12–21 for the binding runs, respectively. Note that different distances are shown in Supplementary Figs. 3–21 to illustrate salt bridges.

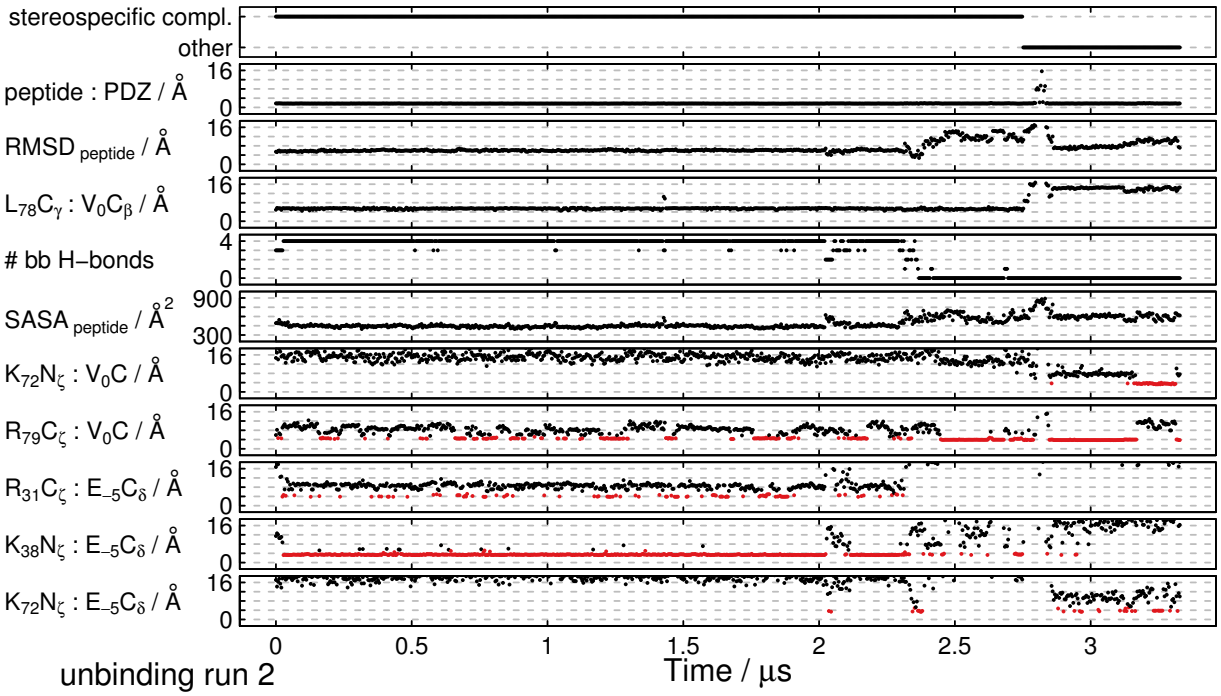


Fig. S3 Analysis of the second unbinding run. Similar to Supplementary Fig. 2.

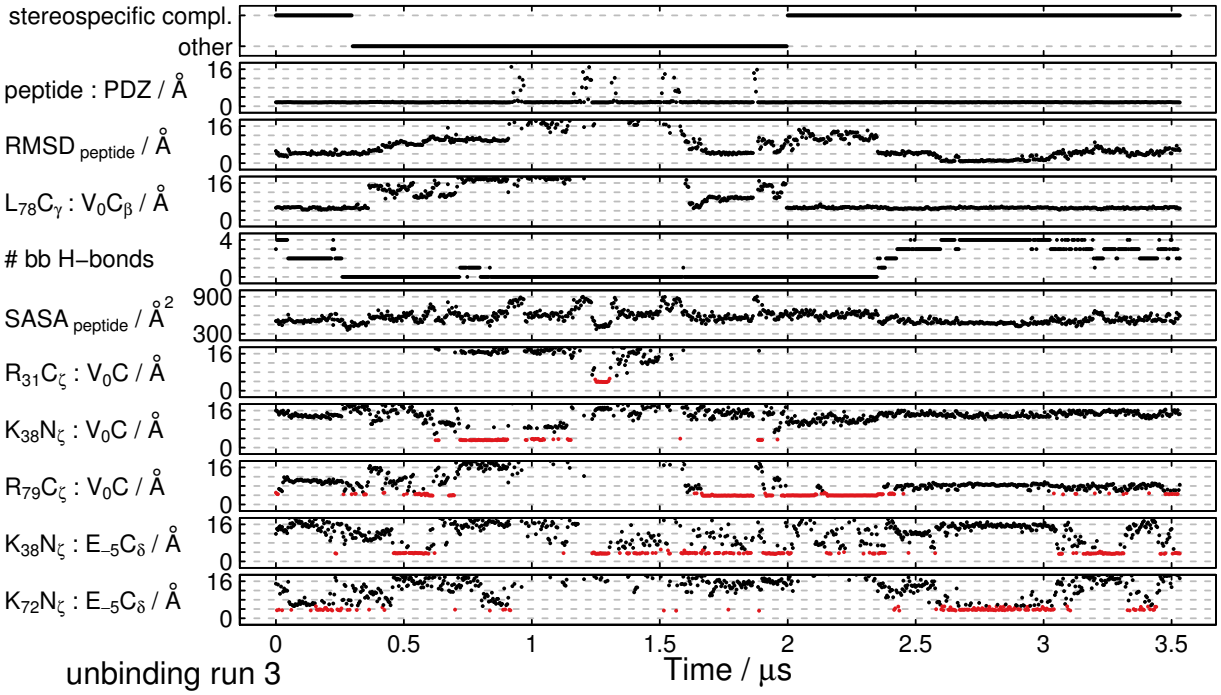


Fig. S4 Analysis of the third unbinding run. Similar to Supplementary Fig. 2.

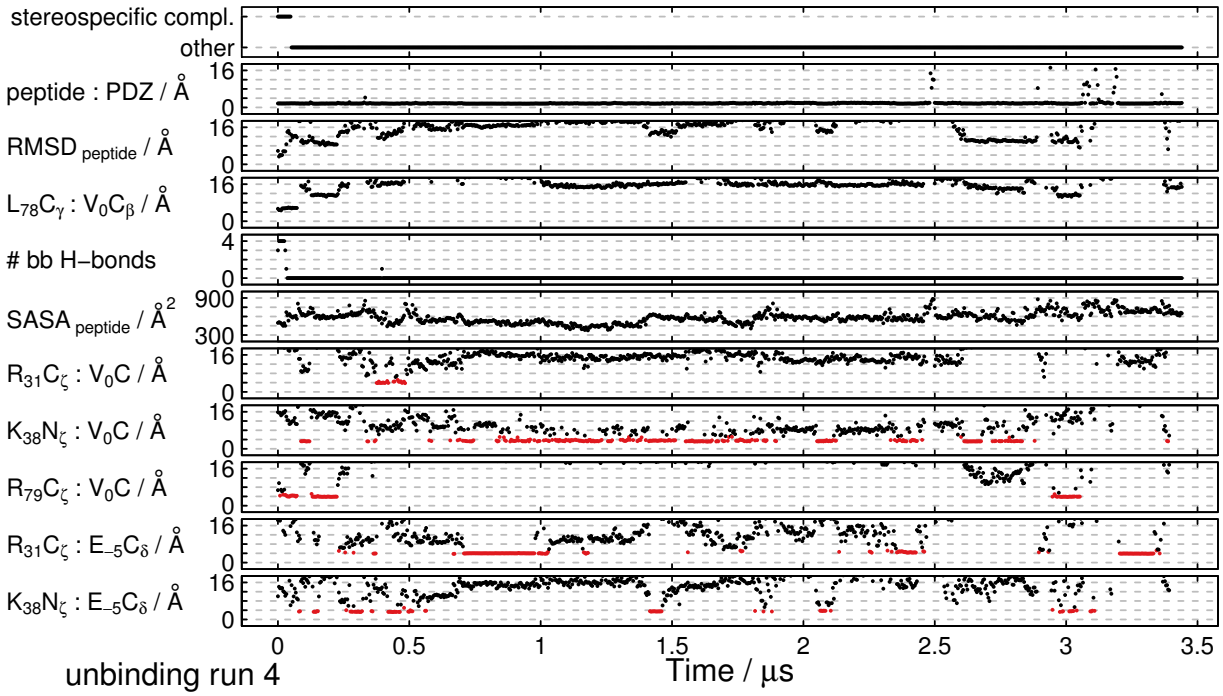


Fig. S5 Analysis of the fourth unbinding run. Similar to Supplementary Fig. 2.

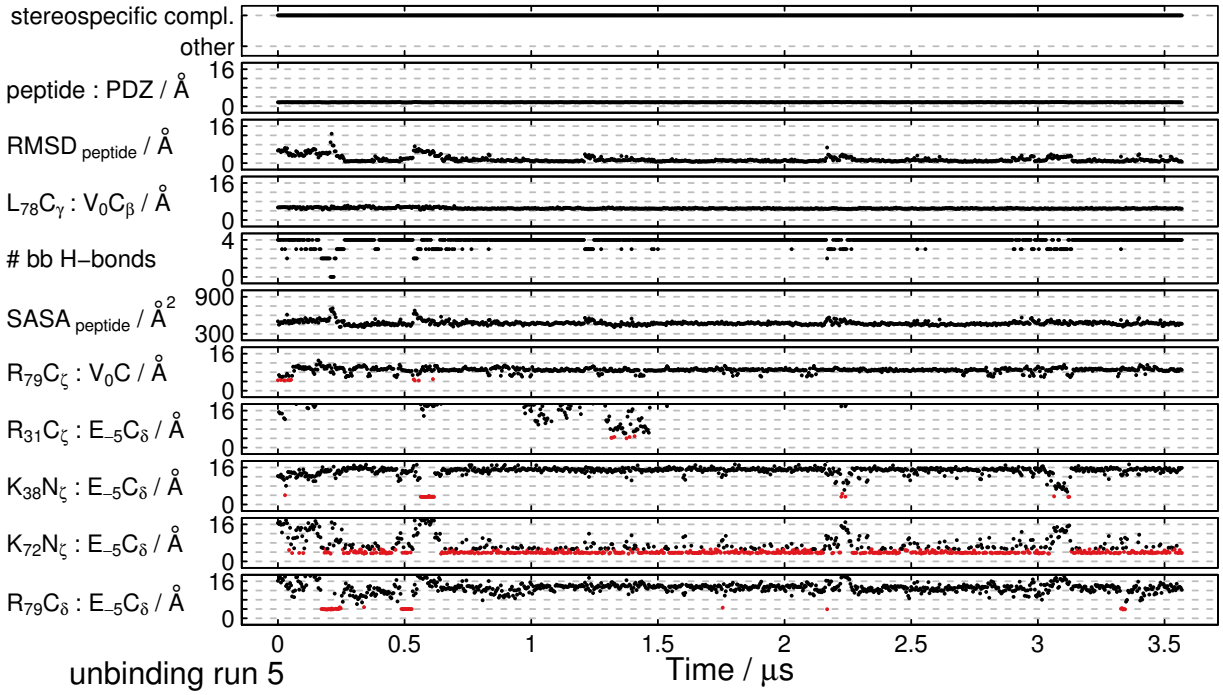


Fig. S6 Analysis of the fifth unbinding run. Similar to Supplementary Fig. 2.

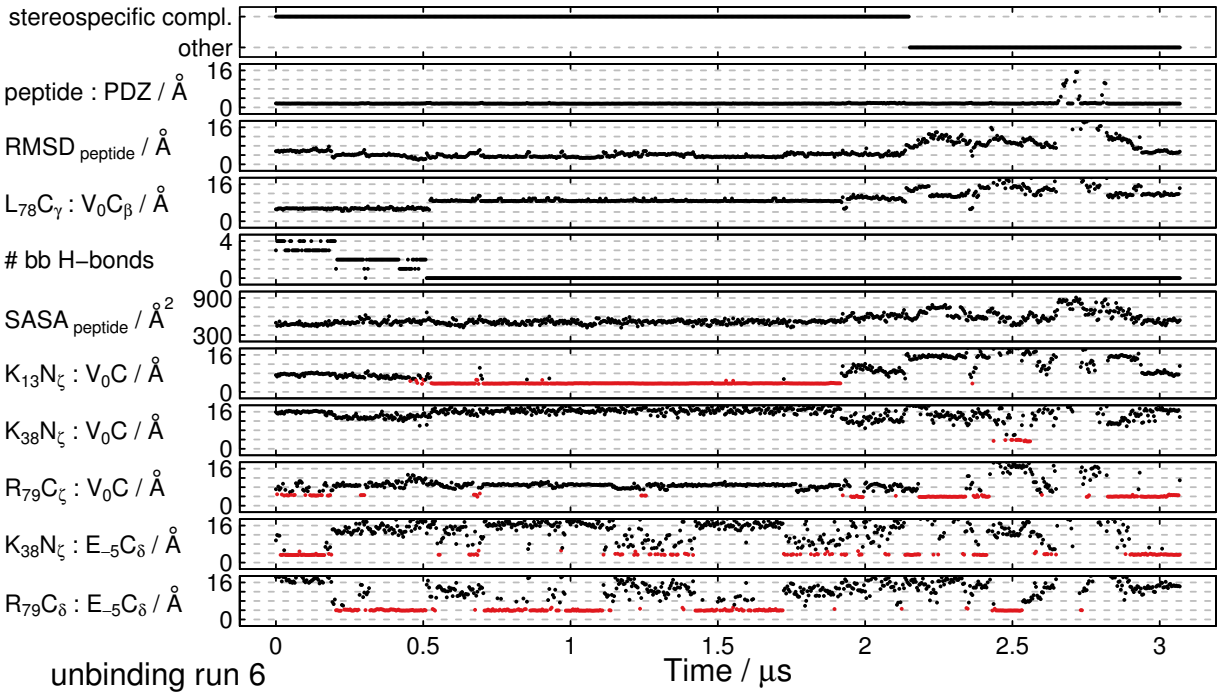


Fig. S7 Analysis of the sixth unbinding run. Similar to Supplementary Fig. 2.

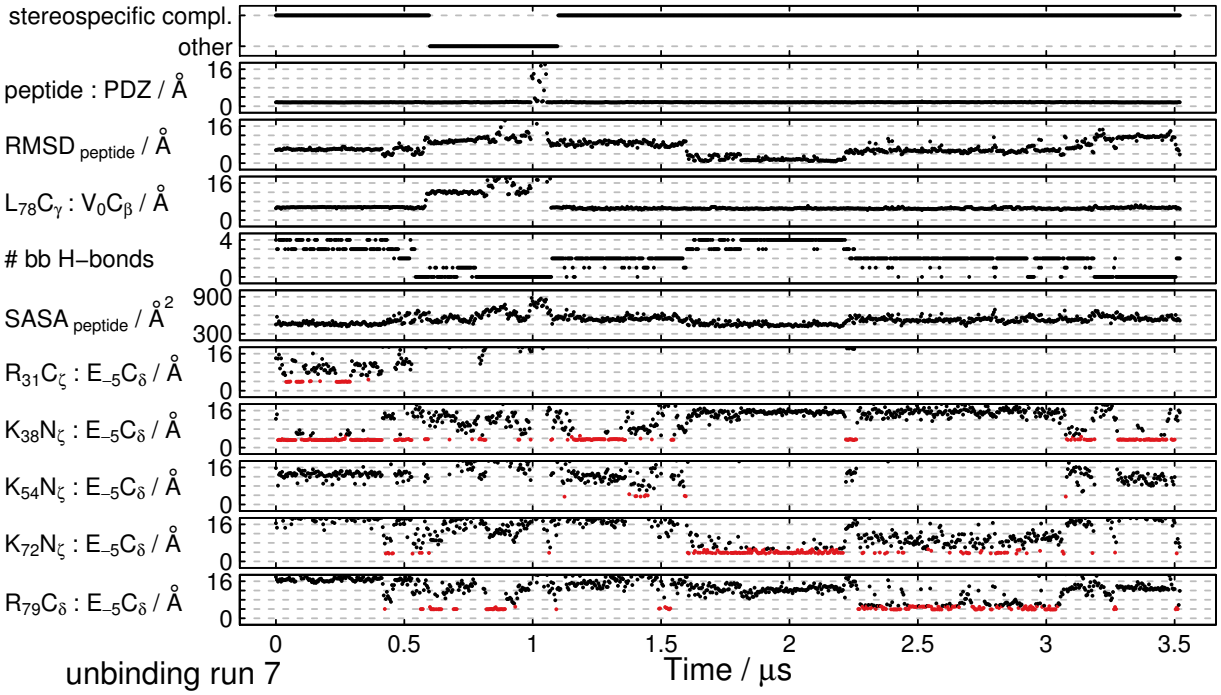


Fig. S8 Analysis of the seventh unbinding run. Similar to Supplementary Fig. 2.

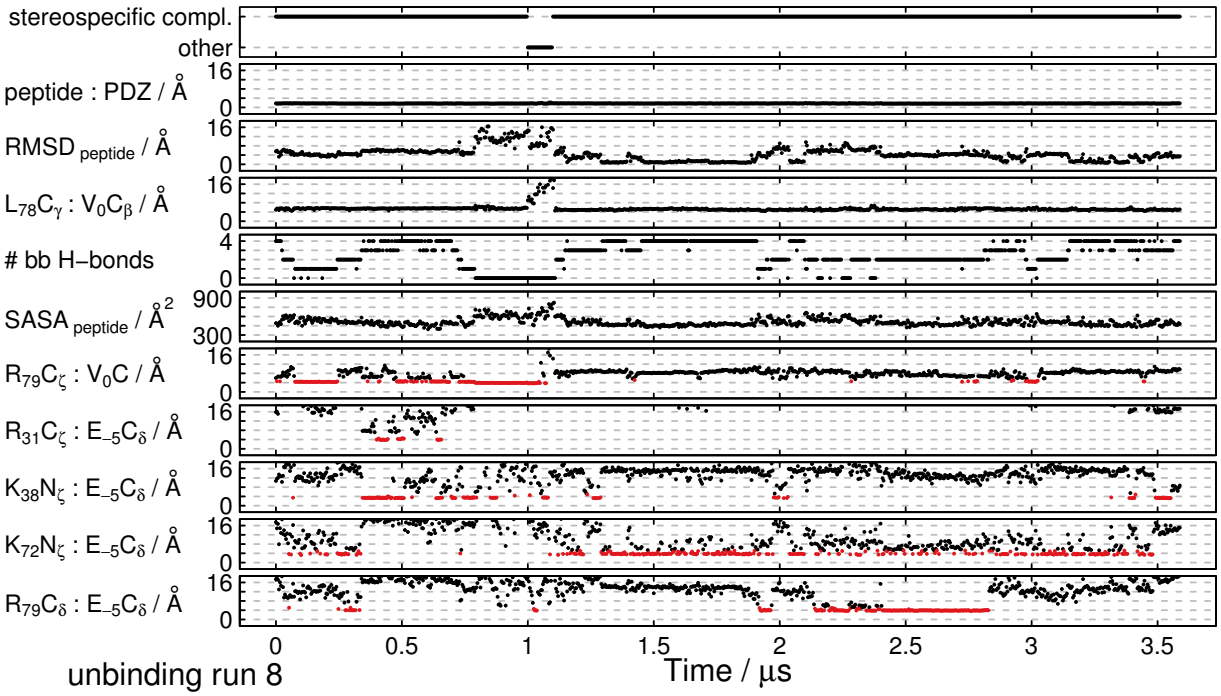


Fig. S9 Analysis of the eighth unbinding run. Similar to Supplementary Fig. 2.

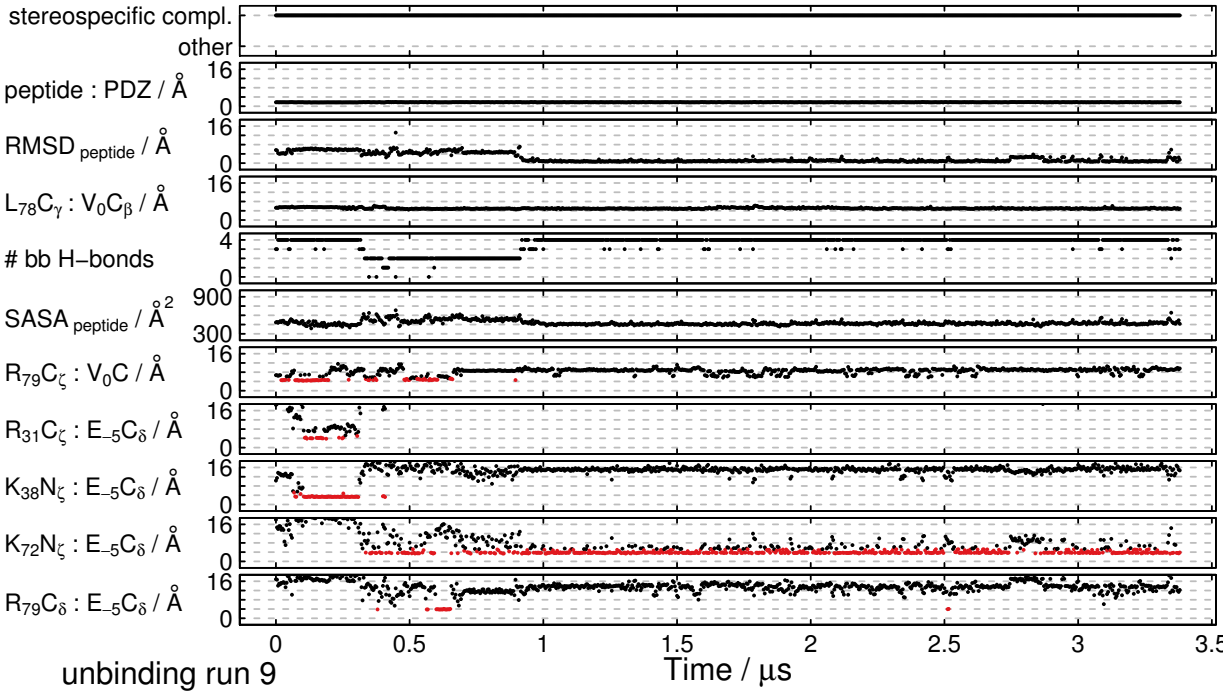


Fig. S10 Analysis of the ninth unbinding run. Similar to Supplementary Fig. 2.

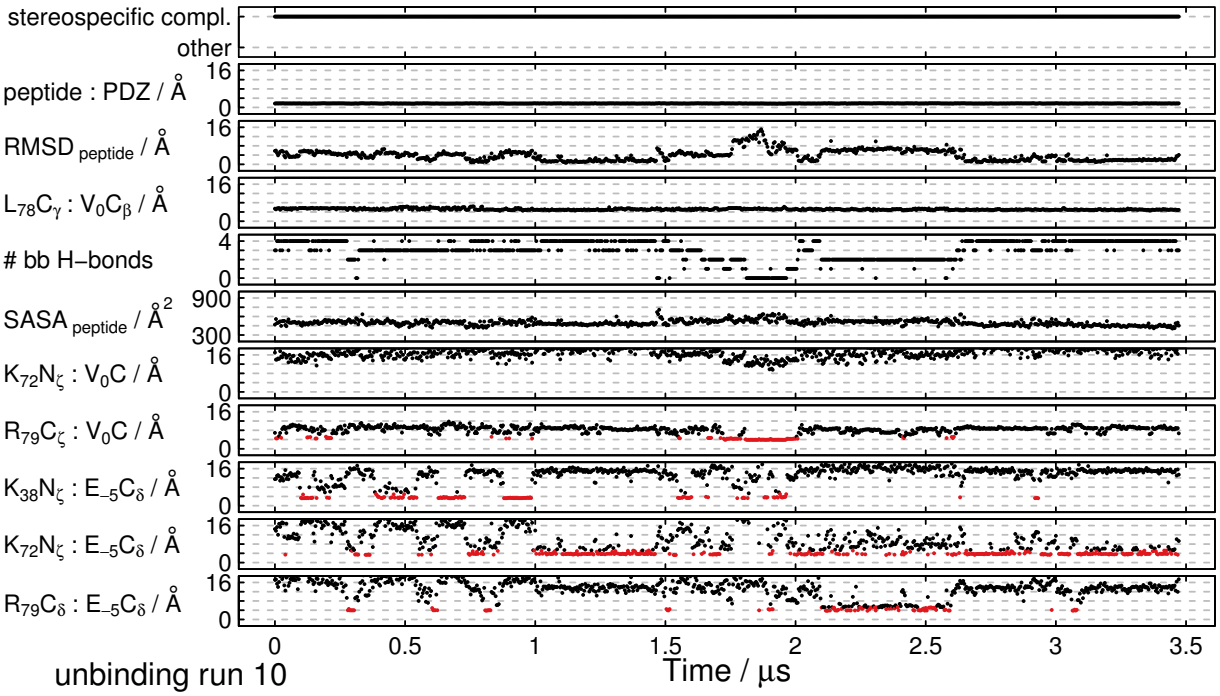


Fig. S11 Analysis of the tenth unbinding run. Similar to Supplementary Fig. 2.

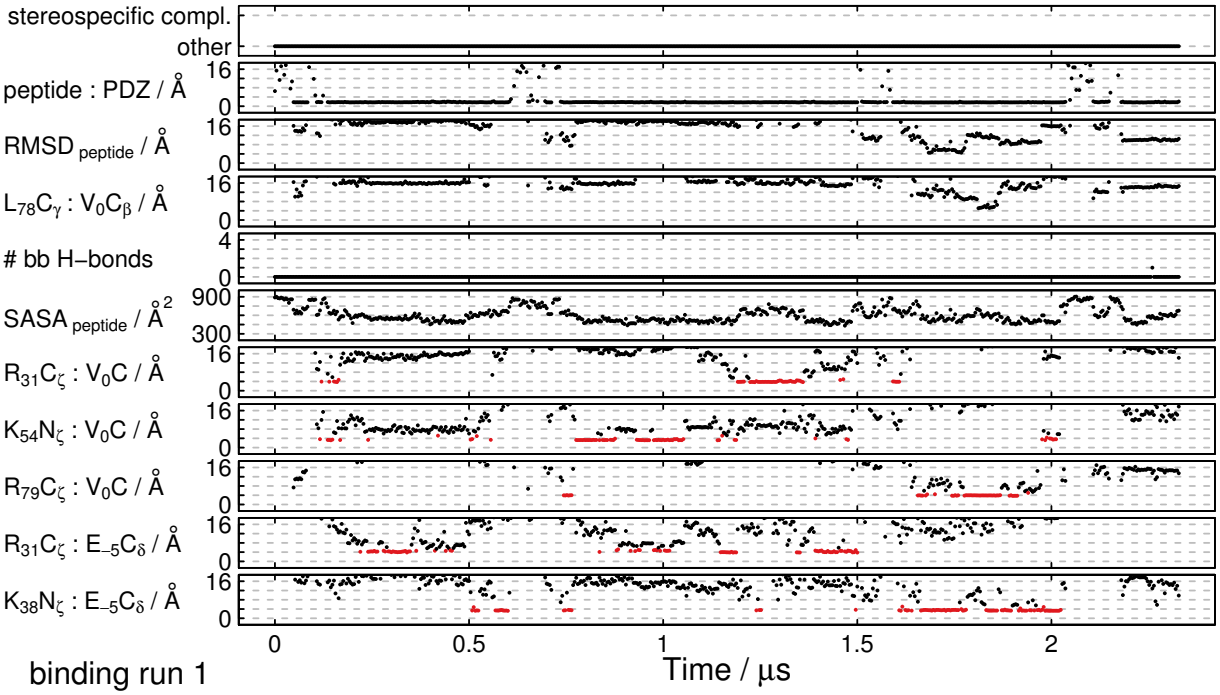


Fig. S12 Analysis of the first binding run. Similar to Supplementary Fig. 2.

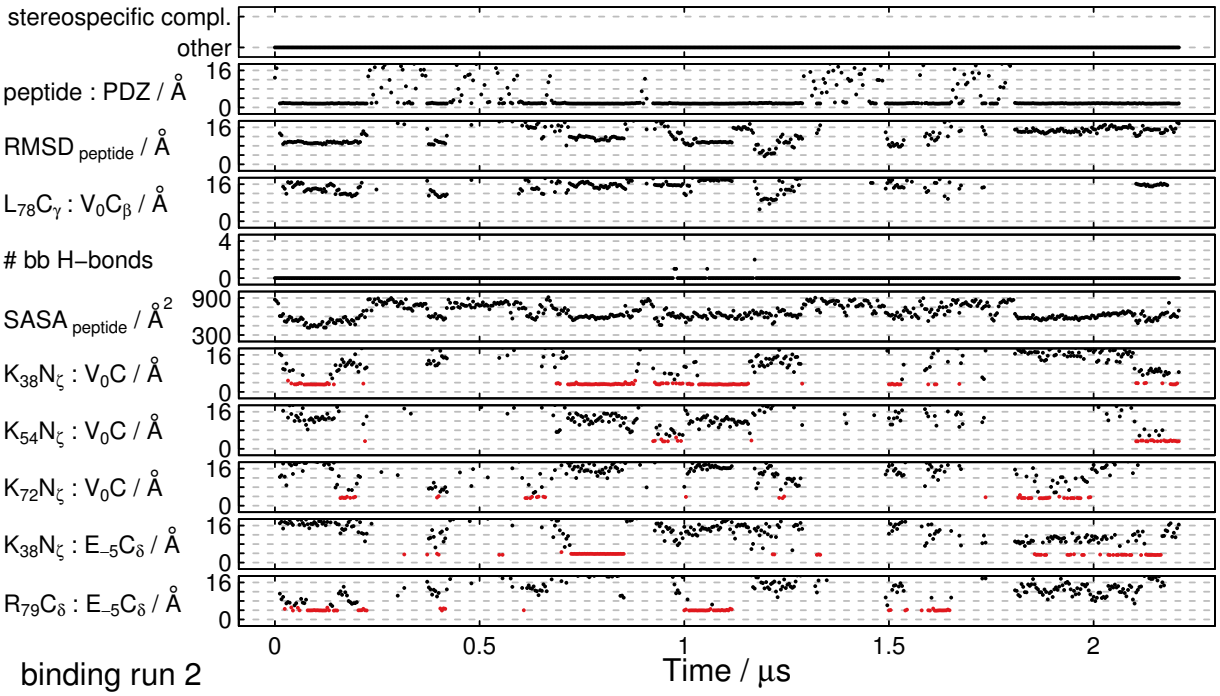


Fig. S13 Analysis of the second binding run. Similar to Supplementary Fig. 2.

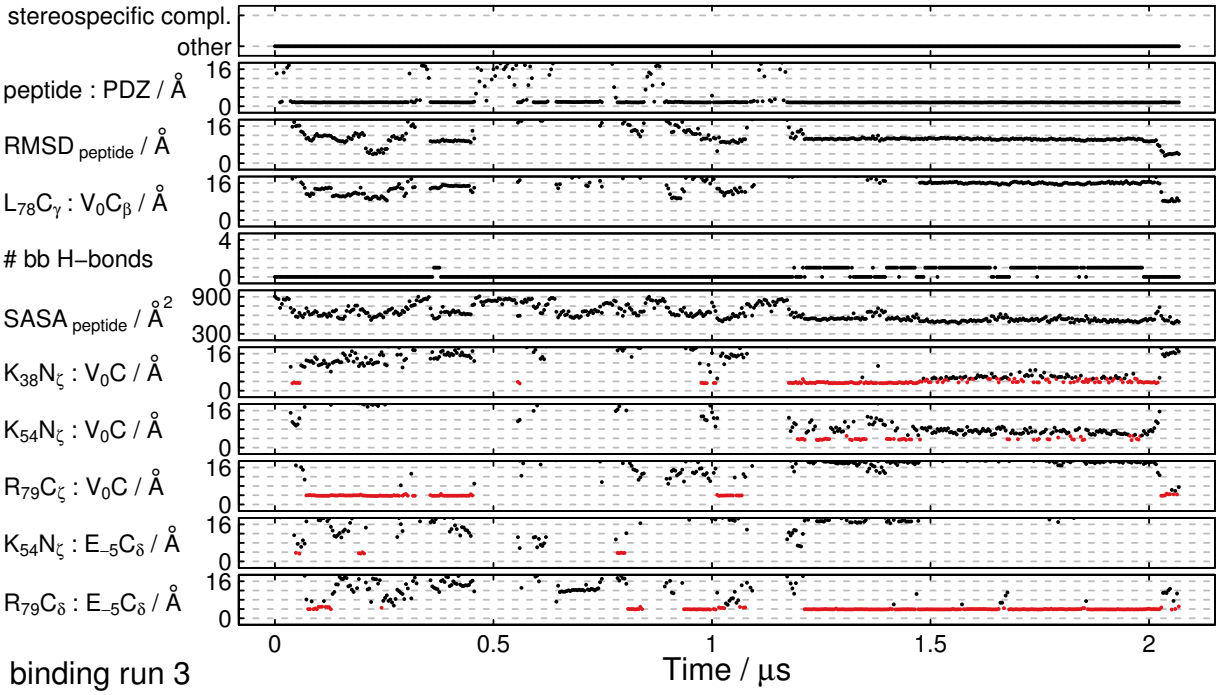


Fig. S14 Analysis of the third binding run. Similar to Supplementary Fig. 2.

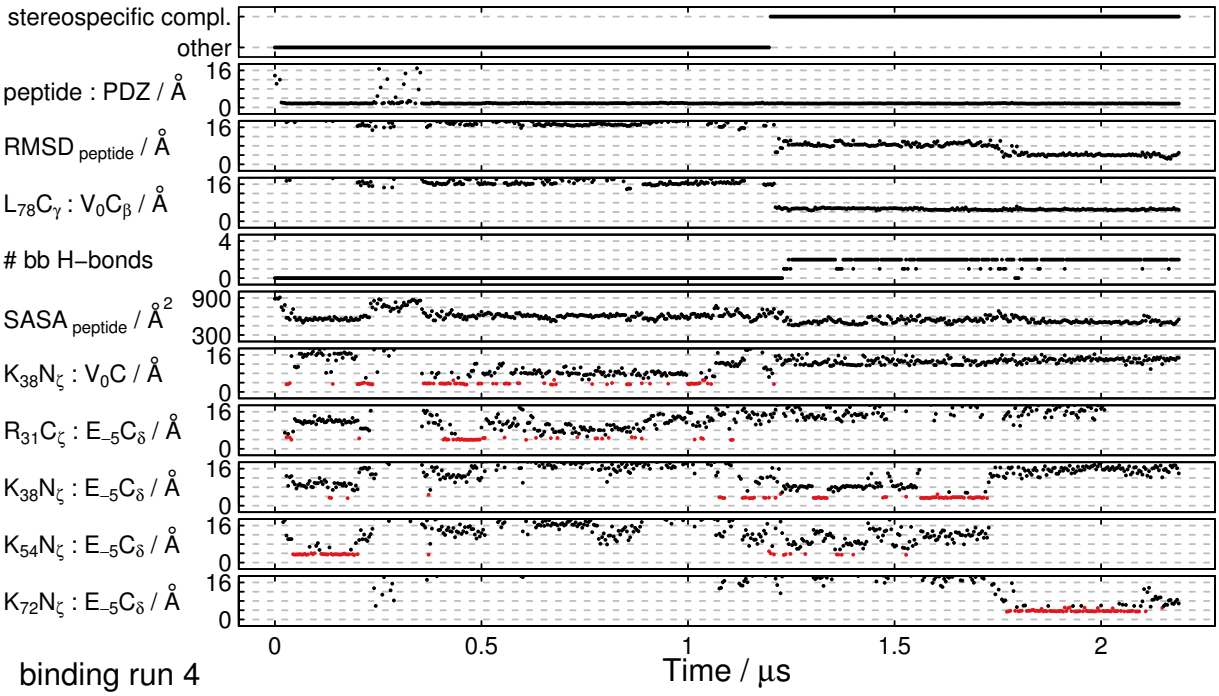


Fig. S15 Analysis of the fourth binding run. Similar to Supplementary Fig. 2.

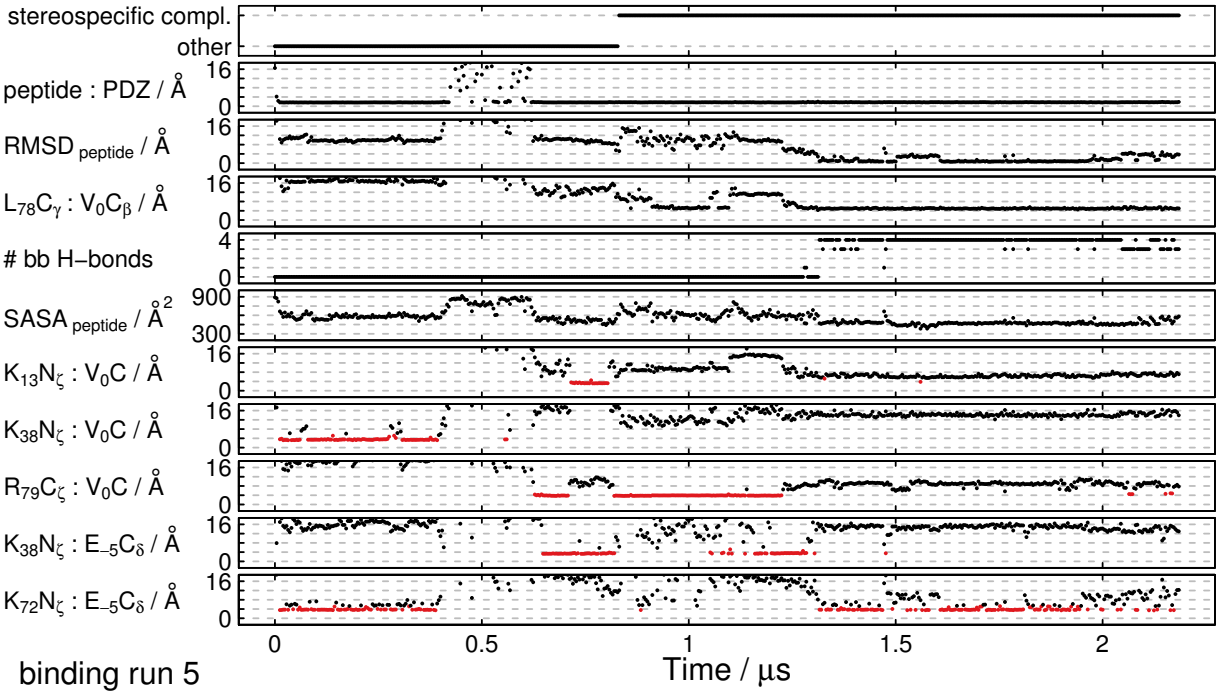


Fig. S16 Analysis of the fifth binding run. Similar to Supplementary Fig. 2.

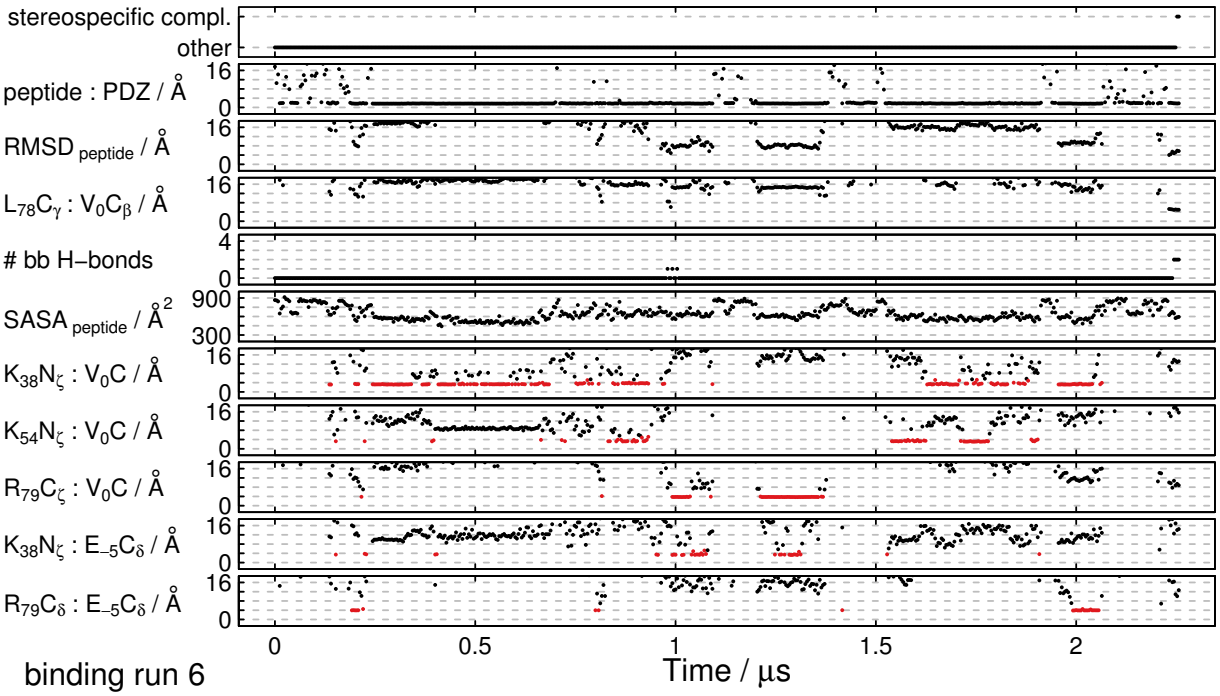


Fig. S17 Analysis of the sixth binding run. Similar to Supplementary Fig. 2.

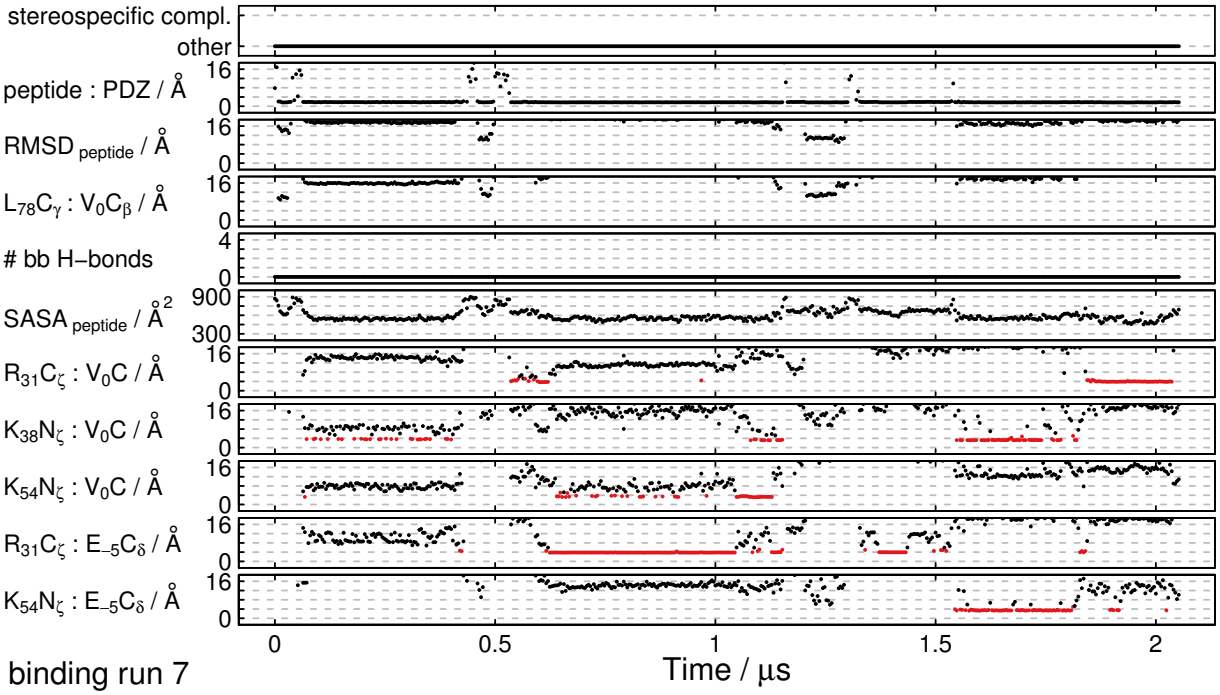


Fig. S18 Analysis of the seventh binding run. Similar to Supplementary Fig. 2.

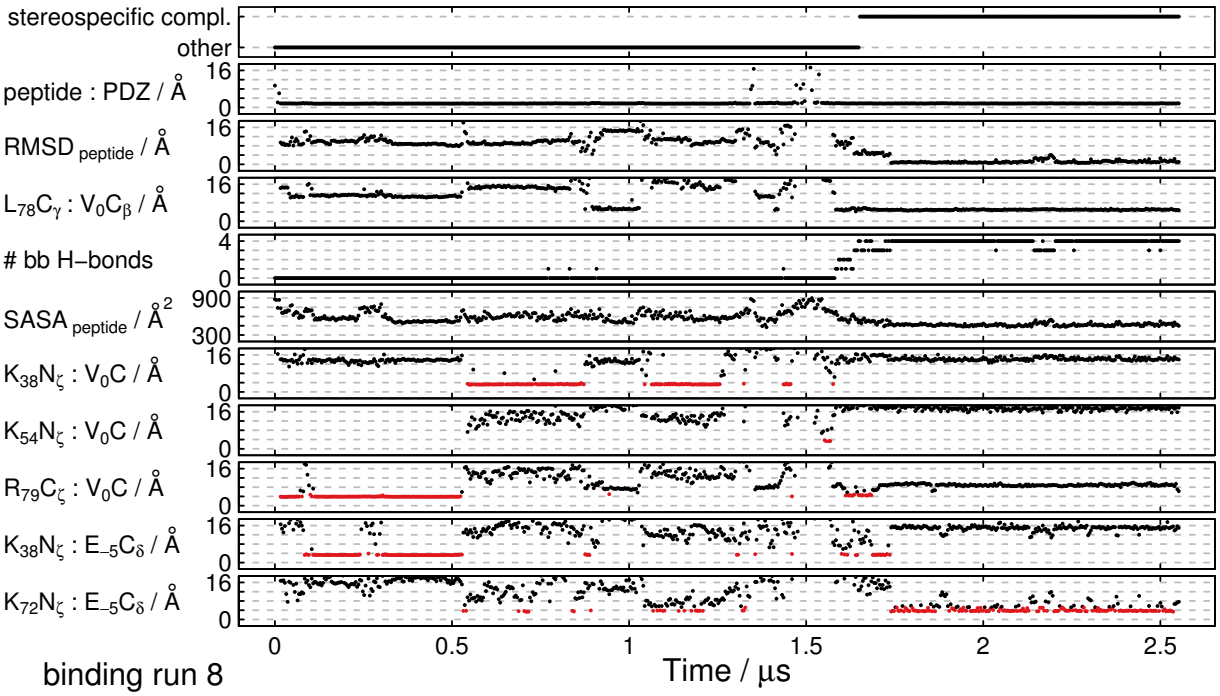


Fig. S19 Analysis of the eighth binding run. Similar to Supplementary Fig. 2.

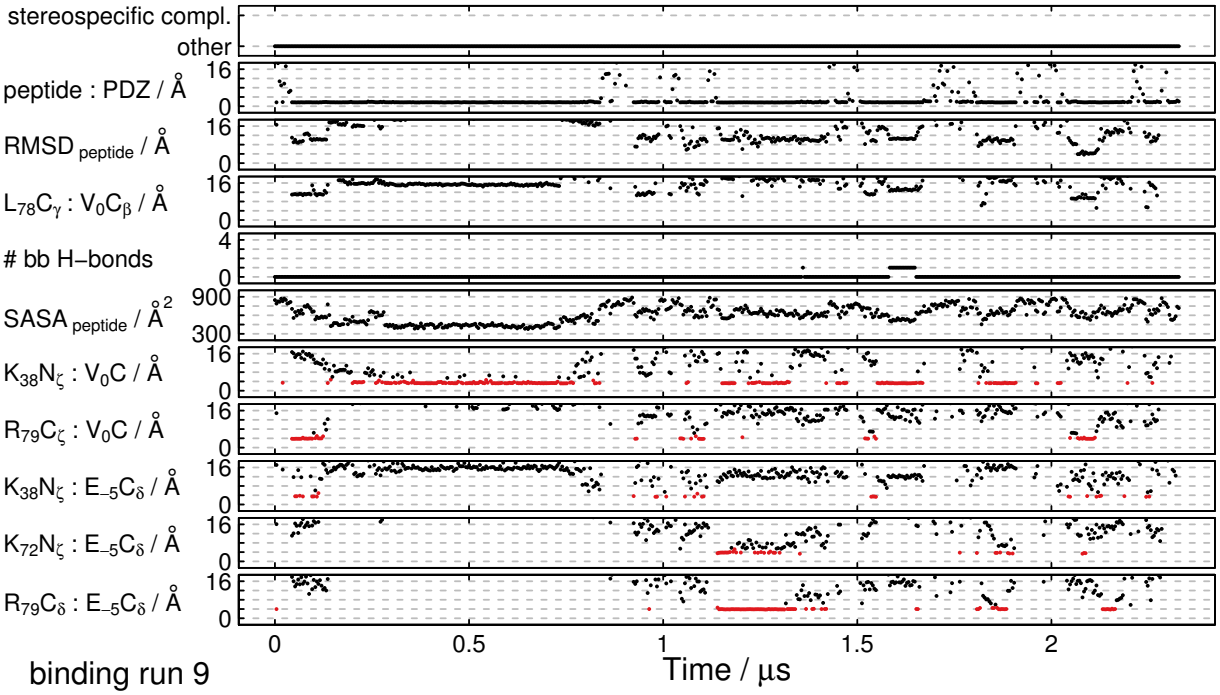


Fig. S20 Analysis of the ninth binding run. Similar to Supplementary Fig. 2.

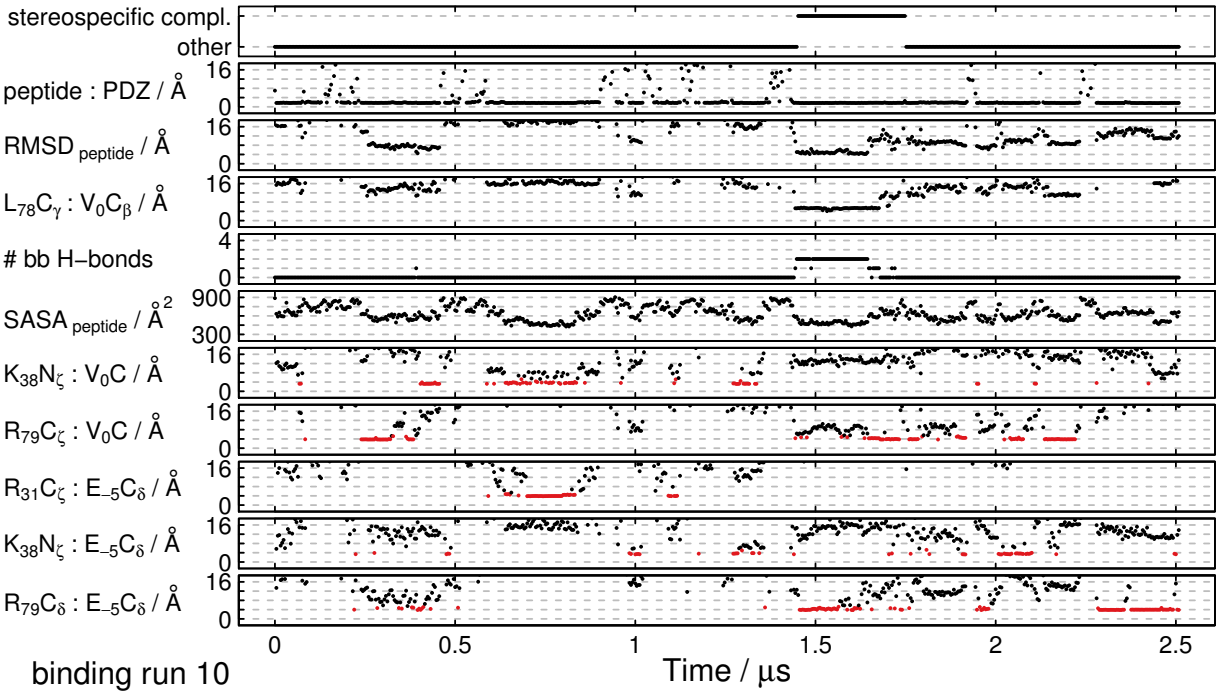


Fig. S21 Analysis of the tenth binding run. Similar to Supplementary Fig. 2.

Group	Atom pair			
carboxylate-binding loop	Ser17	H_γ	Val0	C
	Leu18	H_N	Val0	C
	Gly19	H_N	Val0	C
	Ile20	H_N	Val0	C
β2	Ile20	O	Val0	H_N
	Val22	H_N	Ser-2	O
	Val22	O	Ser-2	H_N
	Gly24	H_N	Gln-4	O
α2	His71	N_ϵ	Ser-2	H_γ
	Arg79	C_ζ	Ala-1	O
	Arg79	C_ζ	Val0	C
hydrophobic pocket	Leu18	C_β	Val0	C_β
	Leu18	C_γ	Val0	C_β
	Ile20	$C_{\gamma,1}$	Val0	C_β
	Ile20	C_δ	Val0	C_β
	Val22	C_β	Val0	C_β
	Val75	C_β	Val0	C_β
	Leu78	C_β	Val0	C_β
	Leu78	C_γ	Val0	C_β
	Arg79	C_γ	Val0	C_β
N-terminus of ligand	Thr23	C_α	Glu-5	C_α
	Thr23	C_α	Gln-4	C_α
	Thr23	C_α	Gln-4	C_δ
	Thr23	C_α	Val-3	C_α
	His71	C_α	Glu-5	C_α
	His71	C_α	Gln-4	C_α
	His71	C_α	Gln-4	C_δ
	His71	C_α	Val-3	C_α
	Asn27	N_δ	Gln-4	O_ϵ

Table S1 Distance function used for SAPPHIRE plot. This table lists the 29 atom pairs for the distance function used for the SAPPHIRE plot shown in Fig. 3 of the main text. Distances are grouped structurally. The group 'hydrophobic pocket' contains atoms from the hydrophobic pocket surrounding the side chain of Val0 in the crystal structure, and the group 'N-terminus of ligand' is meant to capture the orientation of the N-terminal part of the ligand with respect to the protein. The distance between two snapshots i and j is given by $\sqrt{\sum_{k=1}^{29}(d_k^i - d_k^j)^2}$ where d_k^i is the distance between the k-th atom pair in snapshot i.

Chapter 5

Weighted distance functions improve analysis of high-dimensional data: application to molecular dynamics simulations

Blöchliger, N., Caflisch, A. and Vitalis, A. *Journal of Chemical Theory and Computation*, 11(11): 5481–5492, 2015

JCTC Journal of Chemical Theory and Computation

Article

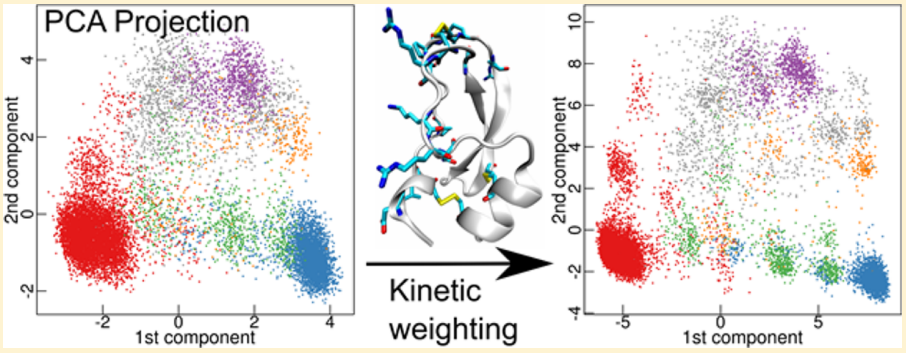
pubs.acs.org/JCTC

Weighted Distance Functions Improve Analysis of High-Dimensional Data: Application to Molecular Dynamics Simulations

Nicolas Blöchliger, Amedeo Caflisch, and Andreas Vitalis*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Zurich, Switzerland

Supporting Information



ABSTRACT: Data mining techniques depend strongly on how the data are represented and how distance between samples is measured. High-dimensional data often contain a large number of irrelevant dimensions (features) for a given query. These features act as noise and obfuscate relevant information. Unsupervised approaches to mine such data require distance measures that can account for feature relevance. Molecular dynamics simulations produce high-dimensional data sets describing molecules observed in time. Here, we propose to globally or locally weight simulation features based on effective rates. This emphasizes, in a data-driven manner, slow degrees of freedom that often report on the metastable states sampled by the molecular system. We couple this idea to several unsupervised learning protocols. Our approach unmasks slow side chain dynamics within the native state of a miniprotein and reveals additional metastable conformations of a protein. The approach can be combined with most algorithms for clustering or dimensionality reduction.

1. INTRODUCTION

The analysis of high-dimensional data is susceptible to several pitfalls.[1−4] Most unsupervised learning methods, such as clustering or dimensionality reduction, require a notion of similarity or distance between individual observations or snapshots. If individual snapshots are vectors of high dimensionality, most functional forms measuring distance lack contrast, *i.e.*, for a given query point the nearest and farthest data points are almost equally far from it.[5,6] Additional problems arise because the data might contain a large number of irrelevant features (dimensions), and because the importance of features can differ for different data points or clusters.[7−9] As a consequence, the choice of a distance function offering sufficient contrast can be more important than the choice of learning method.[10−12] This calls for efficient protocols to derive similarity measures that do not suffer from lack of contrast and account for local feature relevance. These measures should be accessible without an intricate understanding of the system described by the data.

For high-dimensional data, it is common to select or generate features that are deemed informative. When performed manually, this process relies primarily on domain expertise. Measures of relevance, such as entropy or mutual information, can serve as guides to nonexpert users.[13] The term feature extraction is commonly associated with techniques of dimensionality reduction.[13] Many of these techniques try to generate new features that maximize a target property, *e.g.*, variance in principal component analysis.[14] Low-dimensional embeddings of high-dimensional data might be of limited use if these data contain many irrelevant features, and if the chosen distance function is unable to distinguish between similar and dissimilar points. It has been noted that feature selection prior to dimensionality reduction can improve the discriminatory power of the latter.[15] Lastly, the contrast level offered by a given distance function may also depend on the position of the two points in data space, and this is reflected in clustering algorithms with locally adaptive similarity measures.[8,9]

Here we focus on high-dimensional data from molecular dynamics (MD) simulations of biomolecules.[16] At its core, analysis of MD data is often concerned with identifying metastable conformations of the simulated system.[17−19] Unsupervised learning methods for this purpose include clustering and related techniques,[11,20−25] classical dimension-

Received: June 30, 2015
Published: October 27, 2015

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

ality reduction algorithms and modifications thereof,[14,26−33] as well as other approaches.[34,35] The success of all these methods depends on the careful selection of features or an informative distance function; however, a lot of trial-and-error is used in practice to improve results.

In this contribution, we present an efficient method to either globally or locally weight features according to a notion of relevance. Recognizing that features exhibiting slow modes are more likely to report on metastable states, we define weights based on effective rates. Global weights employ the autocorrelation function, while locally adaptive weights are a function of transition rates within a time window along the trajectory. We apply these approaches to an illustrative model system and two data sets generated by MD simulations. The first set of MD data originates from simulations of the reversible folding of Beta3S,[36] a 20-residue peptide adopting a three-stranded, antiparallel β-sheet fold. The second example is a very long explicit solvent simulation of the conformational dynamics of bovine pancreatic trypsin inhibitor (BPTI) within its native state.[37] Throughout, we discuss problems that can occur in conjunction with unmodified distance functions and show how weights address them. Where possible, we compare our results to analyses of the same data found in the literature. We show that a comprehensive description of the free energy surface can be extracted from MD trajectories of proteins by including degrees of freedom such as side chains, flexible loops, and terminal residues with appropriate weights. These features are often dismissed *a priori* as noisy and uninteresting, which entails the risk of losing important information.

2. METHODS

Weighted Distance Functions. Consider a set of N observations with each observation corresponding to a data vector of length D. The Euclidean distance between two observations $\mathbf{x}(t_k)$ and $\mathbf{x}(t_l)$ gives equal weight to all their D features:

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l))^2 = D^{-1} \sum_{i=1}^{D} (x_i(t_k) - x_i(t_l))^2 \tag{1}$$

Conversely, the information content relevant for a given target application may differ between features. Given a notion of overall relevance expressed in a vector of weights, $\mathbf{w}$, a weighted Euclidean distance can take into account the heterogeneity of the features as follows:

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l))^2 = \left(\sum_{i=1}^{D} w_i\right)^{-1} \sum_{i=1}^{D} w_i (x_i(t_k) - x_i(t_l))^2 \tag{2}$$

The elements of $\mathbf{w}$ used in eq 2 can represent any notion of importance. Here, we quantify the relevance of features by measurements of net rates obtained independently for each of them. Features associated with low rates are interesting as they are likely to report on metastable states.[38,39] It is expected that a subset of features is homogeneous on the same time scale as the life times of these states. In practice, for the weights in eq 2, we set $w_i = \max(R_i(\tau), 0)$, where $R_i(\tau)$ is the autocorrelation function of the i^{th} feature evaluated at a specific time lag τ. Note that this corresponds to scaling the data and is different from altering the metric itself, *e.g.*, by changing the Euclidean (L_2) to a rectilinear (L_1) norm. In the present work, we often use dihedral angles and represent them by sine and cosine terms. Rather than computing separate weights in this case, we simply keep the larger of the two values derived independently as the resultant weight.

Global weights as used in eq 2 cannot reflect that the importance of individual features might depend on where a given observation is situated in the overall data space. We use locally adaptive weights to account for this. Here, the notion of "local" is derived exclusively from proximity in time, which is a limitation. Unfortunately, the autocorrelation function computed over a data window of width Δ becomes misleading if transitions are absent. Instead, locally adaptive weights are derived by counting the number of times a feature crosses its global mean:

$$n_i^k(\Delta) = \sum_{j=k-\Delta/2}^{j=k+\Delta/2} H(-(x_i(t_{j-1}) - \langle x_i\rangle_N)(x_i(t_j) - \langle x_i\rangle_N))$$
$$w_i^k = (\mathrm{n}_i^k(\Delta) + \alpha)^{-1} \tag{3}$$

Here, H denotes the Heaviside step function, and α is a parameter required to be positive. The weights in eq 3 are expected to be low for features that sample unimodal distributions. If a feature differs between states, eq 3 rewards those features with locally small variances. False negatives can be obtained if the global data mean coincides with a specific peak position in a multimodal distribution. Distance is measured as

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l))^2 = \left(\sum_{i=1}^{D} \sqrt{w_i^k w_i^l}\right)^{-1} \sum_{i=1}^{D} \sqrt{w_i^k w_i^l}\,(x_i(t_k) - x_i(t_l))^2 \tag{4}$$

We note that the function d does not necessarily satisfy the triangle inequality, *i.e.*, eq 4 no longer represents a metric. This may be undesirable. In the context of clustering algorithms, we might also require a measure of distance between an individual observation, $\mathbf{x}(t_k)$, and a group of observations (cluster). Representing the cluster by its unscaled centroid, $\mathbf{c}$, we have

$$d(\mathbf{x}(t_k), \mathbf{c})^2 = \left(\sum_{i=1}^{D} \sqrt{w_i^k w_i^c}\right)^{-1} \sum_{i=1}^{D} \sqrt{w_i^k w_i^c}\,(x_i(t_k) - c_i)^2 \tag{5}$$

In eq 5, $\mathbf{w}^c$ is the average weight vector across all observations that are part of the cluster with centroid $\mathbf{c}$.

Progress Index and SAPPHIRE Plots. Recently, we have developed an algorithm for the analysis of long MD trajectories.[34,35] The resulting SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot is a comprehensive visualization of the thermodynamics and kinetics of the simulated system and is used here to study the performance of the distance functions introduced above.

We briefly describe the method next and refer the reader to the original publications for more details.[34,35] Specifically, all snapshots are assumed to form a complete graph, and the minimum spanning tree or an approximation to it is computed. From a given starting snapshot, the snapshot connected by the shortest available edge is added to a growing partition. The resulting sequence, the so-called progress index, proceeds through regions of high sampling density one after another and avoids overlap of distinct states.[34] The progress index can be annotated to yield a SAPPHIRE plot as described in recent work.[35] Here, we employ the following annotation functions to

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

highlight and interpret the states along the progress index. First, we use a kinetic annotation function to localize the individual states on the progress index. Specifically, for every snapshot i along the progress index, we plot the average of the mean first-passage times between A_i and S_i, denoted τ_{MFP}, where A_i is the set of snapshots added to the progress index before i and S_i is the set of those added after i. The value of this annotation function is low within a state and high in transition regions, and barriers are highlighted reliably (although they cannot be interpreted quantitatively).[34] Second, we plot the actual sampling time of the individual snapshots to illustrate when and in which sequence the different states were sampled. Third, we characterize the states themselves by structural annotations. For Beta3S, we have used the secondary structure assignment according to the DSSP algorithm[40] and the χ_1 angle of Trp10. For BPTI, we show selected dihedral angles using binning with boundaries given in the Supporting Methods. The boundaries were obtained from direct inspection of the individual histograms for each angle. In addition, we show state assignments according to Shaw et al.[37] and Xue et al.[41]

The method is implemented in the CAMPARI simulation and analysis package (http://campari.sourceforge.net). Detailed parameter settings are given in the Supporting Methods. In contrast to previous work, we modify the underlying spanning tree before computing the progress index (Vitalis, manuscript submitted). In particular, we collapse the leaves into their parent vertex, which means that they are added to the progress index as soon as it encounters their parent vertex. This places snapshots from the fringe region around regions of high sampling density next to the snapshots from the closest state. The procedure can be repeated a number of times, and this is a controllable parameter. It is set by CAMPARI keyword FMCSC_CPROGMSTFOLD, which was 1 throughout except for Figure 2 (where it was 2).

Clustering and Cut-Based Free Energy Profiles. Besides SAPPHIRE plots, we employ clustering and cut-based free energy profiles[23] (cfeps) to study the influence of the distance function. Clustering according to a recent, tree-based algorithm partitions the data into tight clusters that have little overlap and are of controllable size.[20] Cfeps order the resultant clusters by their kinetic distance from a chosen reference state. The ordering is kinetically annotated with τ_{MFP}, defined as above. As for SAPPHIRE plots, the value of τ_{MFP} is expected to be low within a basin and high in transition regions. This is what allows an immediate partitioning into metastable states.

3. RESULTS

To illustrate the problems that occur when analyzing data without feature selection, we use a model system and two high-dimensional real-world data sets from MD simulations of the peptide Beta3S[36] and the protein BPTI[37] obtained in implicit and explicit solvent, respectively. We highlight the performance of the different similarity measures by employing a recently developed algorithm for the analysis of dynamical systems that uses a distance function as its only essential parameter.[34] The similarity (or better, dissimilarity) measures evaluated are the unweighted Euclidean distance (UW), the Euclidean distance weighted by the global autocorrelation function at fixed lag time per dimension (GW), and a locally adaptive distance defined by time-local transition rates (LAW). They are defined in eqs 1, 2, and 4, respectively (see Methods). We demonstrate that the weighted distance functions, GW and LAW, offer substantial benefits in all cases investigated. For brevity, we will repeatedly refer to the three dissimilarity measures as UW, GW, and LAW measures below.

Model System. Figure 1a schematically depicts a Markov model of 4 states and its associated transition matrix with the states identified by color throughout. A Markov chain (random walker) is used to generate a continuous trajectory of length 2×10^5 snapshots, which means that even the least likely (red) state is sampled sufficiently. To be able to meaningfully test

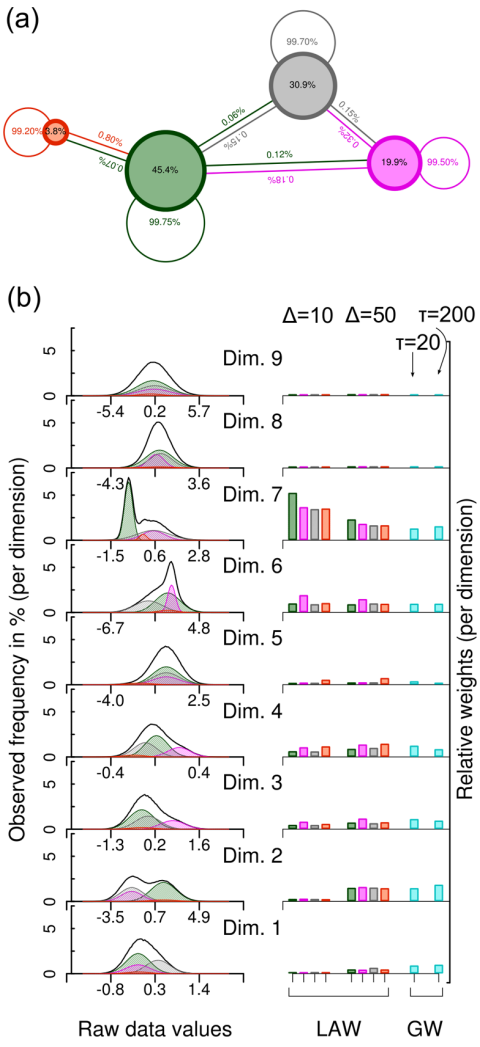


Figure 1. Model system and its representation. (a) Schematic description of the 4-state Markov model. The text within circles gives the steady-state population of each state. Nonzero elements of the transition matrix are shown as lines with the conditional probabilities indicated. Coloring of lines is by source state. (b) Each snapshot of the model is represented by 9 features. Features are generated by independent, memory-free, Gaussian processes with parameters that depend on the macrostate. On the left, we plot actual histograms (black lines) from a trajectory of 2×10^5 snapshots along with the generating functions scaled according to the steady-state population of each state (shaded areas). On the right, weights computed for the same trajectory are shown for each dimension for both LAW and GW measures. Locally adaptive weights are averaged separately for the true state the trajectory resided in and produced for two different window sizes, $\Delta = 10$ and $\Delta = 50$, with $\alpha = 0.01$ (see eq 4). Global weights are computed at two different lag times ($\tau = 20$ and $\tau = 200$).

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

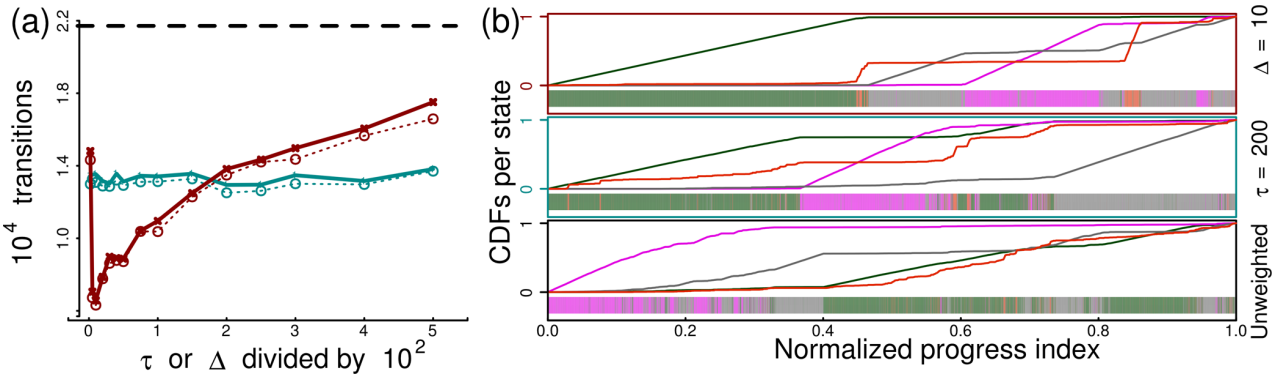


Figure 2. Evaluation of different distance functions for the model system in Figure 1. (a) The number of transitions between states in the progress index is shown. The construction of the progress index relies on preorganization via clustering, and we made use of a recent improvement to the algorithm (see Methods). Each condition for both types of weights was evaluated for 8 (GW, cyan lines) or 5 (LAW, dark red lines) different clustering settings, and the medians (solid lines) and minima (dotted lines) are plotted. The black dashed line is the minimum value for the unweighted case. (b) The exact state annotation (color bar) along the progress index is plotted for every 10th snapshot. Cumulative distribution functions were analyzed and normalized independently for each state. From bottom to top, we show the data for UW, GW, and LAW measures, respectively.

different distance measures for this system, it is represented by 9 data dimensions (features). Every feature is generated from a normal distribution whose parameters depend on the state the system currently resides in. As seen in Figure 1b (left-hand side), no feature is informative for all states. The overlap is generally large, and two features (#8 and #9) are completely uninformative. Despite the moderate dimensionality, this challenges the UW measure.

In Figure 1b we also compare the resultant global and locally adaptive weights underlying the GW and LAW measures, respectively. The global weights obtained from the autocorrelation function at fixed lag time de-emphasize features #5, #8, and #9 irrespective of lag time. At τ = 20, all remaining dimensions have roughly equivalent weights, whereas at τ = 200 features #2 and #7 dominate. These correspond exactly to the histograms with the clearest peak separations. Since we know the correct state for each snapshot, the locally adaptive weights can be averaged separately for different states. These weights correspond to the inverse crossing rate of the global data mean for a given feature (eq 4). This is why they emphasize features that have low variance for a given state, *e.g.*, #6 is particularly important for the magenta state or #7 for the green state. Similarly, they also reflect whether a feature's value in a given state is far away from the global mean, *e.g.*, #5 is only relevant for the red state. Note that these synthetic data are memory-free, *i.e.*, time correlation comes exclusively from state persistence.

We scanned a wide range of possible lag times and window sizes, and the particular values shown in Figure 1b correspond to the top performing cases in the subsequent analysis, which was performed as follows. Using a recent algorithm,[34] we computed the progress index that corresponds to stepping through an approximation of the minimum spanning tree (see Methods for details). This procedure is very sensitive to the distance function in use. Ideally, it should arrange snapshots exactly by their underlying states assuming they are geometrically separable. The large overlap seen in Figure 1b makes this task challenging. As a measure of sorting quality, we simply count the number of times the state annotation changes in the progress index, and these data are shown Figure 2a (lower is better). It is clear that the UW measure is rigorously outperformed by both the GW and LAW measures irrespective of parameter settings. GW is inferior to LAW (in terms of peak performance), and its performance appears to change relatively little with lag time. The results for the LAW measure show a clear preference for window sizes that are considerably less than the average life times of states, which is ∼300 steps. We picked the respective top-performing cases for the results obtained with the UW, GW, and LAW measures and visualize the progress index in Figure 2b. Cumulative distribution functions resolved by state highlight the successive decrease in state overlap when changing from UW (bottom row) to GW (middle row) and finally to LAW (top row) measures. We note the improvement of the localization of the red state in particular. With the UW measure, this state would certainly not have been identified as a metastable state of the system.

Beta3S. The first MD data set is taken from an implicit solvent simulation of the 20-residue antiparallel β-sheet peptide Beta3S.[36] Multiple folding and unfolding events are observed during the total sampling time of 20 μs. The unfolded state ensemble is characterized by the presence of several metastable states that are enthalpically stabilized. The data set consists of 10^6 snapshots saved at an interval of 20 ps. We represent the peptide via 99 dihedral angles. The rotation of 2-fold or 3-fold symmetric groups consisting entirely of hydrogen atoms and χ_2 and χ_3 angles of tyrosine were ignored. Dihedral angles enter as their sine and cosine values to avoid intricacies with circular variables.[26]

First, we investigate how the use of locally adaptive weights affects clustering. We clustered the data according to both UW and LAW measures using a recent, tree-based algorithm[20] with thresholds that yield a total number of clusters within 2% of one another. We identified two clusters in the region of highest sampling density (*i.e.*, in the native state) sharing the same centroid. For adjacent lower density regions (see Figure S1 in Supporting Information for details), we picked two clusters whose centroids differ but which are of similar size and distance from the native state. Because distance functions based on dihedral angles are putatively uninformative, we crosscheck cluster definition against the most common and intuitive distance function, *viz.*, the root-mean-square deviation (RMSD) computed over the Cartesian coordinates of all atoms after

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

pairwise alignment. For the native state, Figure 3 (left) reveals that the quality of the clusters obtained by both UW and LAW

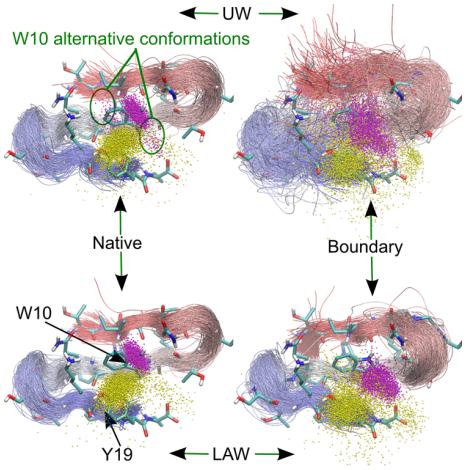


Figure 3. Illustrations of clusters corresponding to the native state and a lower density region, respectively. The clustering uses 99 nonsymmetric dihedral angles in conjunction with UW (eq 1) and LAW (eq 4, Δ = 2 ns, α = 1) measures. Further statistics are provided in the caption of Figure S1. For both measures, we identified clusters in the native basin and near its boundary. For the native state, two clusters could be obtained from UW (35215 snapshots) and LAW (33445 snapshots) measures, which share their centroid snapshot and overlap to 82% identity. For the boundary case, the two clusters shown were identified with the help of Figure S1 (UW: 4013 snapshots; LAW: 3883 snapshots). All cluster members were aligned to the native state centroid (displayed as sticks). For each case, ~500 snapshots are shown in ribbon representation (N-terminus is red). Magenta and yellow spheres document the positions of the NE1 and OH atoms of Trp10 and Tyr19, respectively. These data are shown for ~4000 cluster members. All graphics were rendered with VMD.[42]

measures is comparable, which indicates that excellent sampling density may overcome weaknesses of the distance function. However, more overlaps of alternative conformations are obtained when omitting weights (recognizable most clearly for Trp10). This is confirmed by the RMSD histograms in Figure 4a that exhibit a distinct tail for the cluster based on the UW measure. Such a tail is absent when inspecting the histograms for the UW measure directly (Figure 4b).

We next focus on a region of lower sampling density, *viz.*, clusters situated in the boundary region of the native state (see Figure S1). Figure 3 (top right) demonstrates that the UW measure fails to produce an ensemble satisfying intuitive criteria for what a cluster is. This is rectified by applying the LAW measure, which produces a cluster ensemble that maintains native topology albeit with much increased fluctuations, and that has the side chain of Trp10 in a well-defined region distinct from that of the native state (bottom right). This result is quantified clearly by differences in the histograms of pairwise distances using the RMSD measure (Figure 4c). Obviously, the UW "cluster" contains a wide variety of structures with pairwise distances exceeding 8 Å. There is no difference between self-similarity and similarity to the native state. This is improved dramatically with the LAW cluster, for which the native state clearly is a more dissimilar conformation than the other cluster members.

Figure 4d suggests that the result in Figure 4c for the UW measure is likely due to dimensionality problems, *i.e.*, the

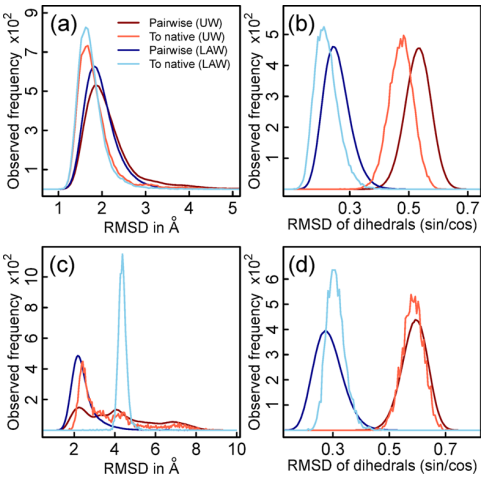


Figure 4. Distance histograms for the clusters in Figure 3. The clustering uses 99 nonsymmetric dihedral angles in conjunction with UW (eq 1) and LAW (eq 4, Δ = 2 ns, α = 1) measures. The legend in panel (a) applies to all panels. (a) For native-state clusters, a total of ~2×10^6 randomly selected and unique pairwise distances of the all-atom coordinate RMSD were computed, and histograms are shown along with complete histograms of distances to the native state centroid (bin size of 0.05 Å). (b) The same as (a) but using the actual UW and LAW measures to compute distances. (c) The same as (a) for the clusters from the boundary region of the native state. (d) The same as (c) but using the actual UW and LAW measures to compute distances.

distance distribution to a snapshot not part of the cluster is almost the same as the pairwise distribution within the cluster. This lack of contrast also holds when analyzing the distance distribution of all snapshots with respect to the native state. In fact, Figure 5a shows that the distribution remains nearly

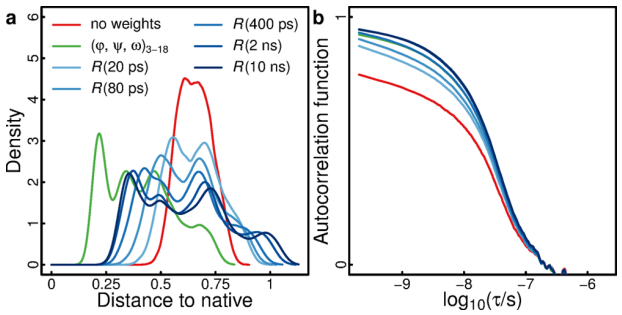


Figure 5. Weighted distance functions capture thermodynamics and kinetics of Beta3S better than an unweighted one. (a) Distance distributions for Beta3S with respect to a representative snapshot of the native basin. Using 103 dihedral angles (including χ_2 and χ_3 angles of tyrosine) and the UW measure, the distribution is essentially unimodal (red curve). With manual feature selection, the distribution has several distinct peaks that indicate coarse clusters in the data (green curve). Here we used the backbone dihedral angles of residues 3–18 as in previous work.[20] Increasing the time lag τ for the GW measure in this range leads to more and better separated peaks (blue curves). The GW measure with a time lag of τ = 2 ns and the Euclidean measure with manual feature selection are correlated (Pearson's correlation coefficient ρ = 0.978). (b) Autocorrelation functions of the distance time series used in (a). For this figure, distances were only computed for every 10th snapshot in the trajectory.

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481–5492

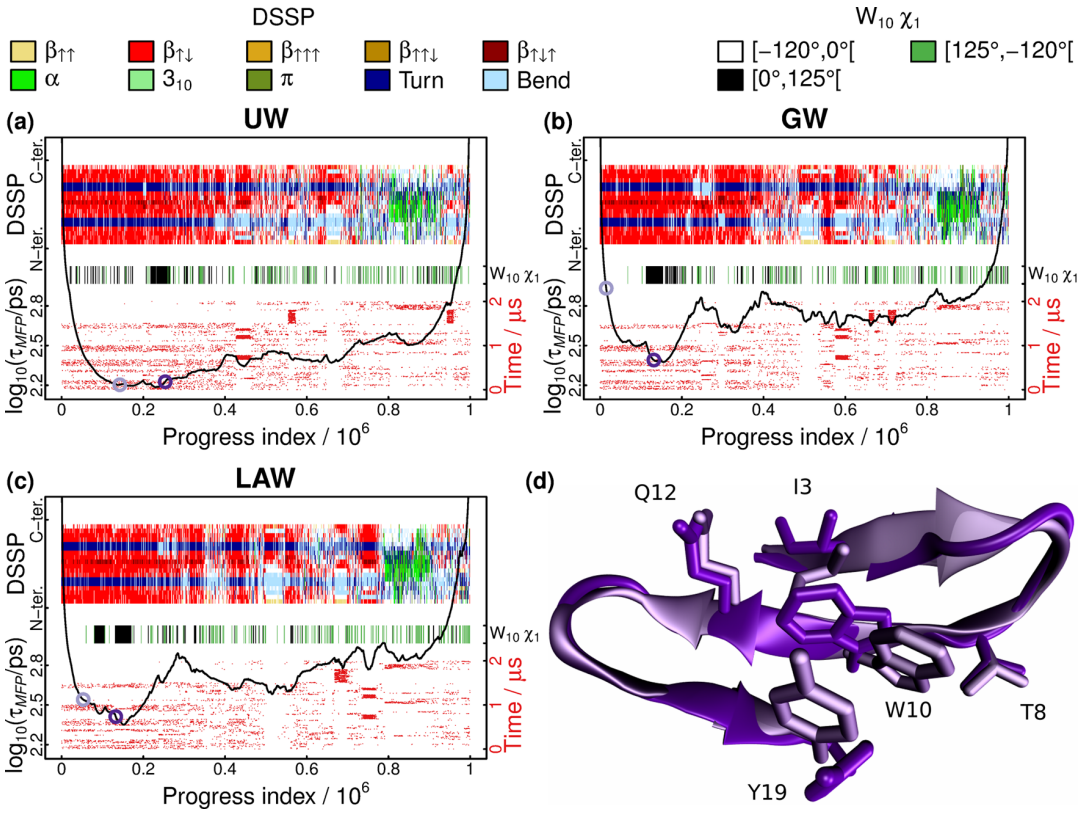


Figure 6. SAPPHIRE plots for Beta3S. (a) SAPPHIRE plot for Beta3S obtained with the UW measure. The peptide is represented by the sine and cosine values of 99 nonsymmetric dihedral angles. The progress index (x axis) represents a reordering of the trajectory snapshots that groups similar snapshots next to each other (see Methods). It is annotated with kinetic information (τ_{MFP}, a function whose value is low within states and high in transition regions, black profile in the bottom), sampling time (red dots, only shown for one out of 10 simulation runs), DSSP assignment[40] by residue (legend on top), and the χ_1 angle of Trp10 (legend on top). (b) The same as (a) for the GW measure with τ = 2 ns. (c) The same as (a) for the LAW measure with Δ = 2 ns and α = 1. All profiles in (a)−(c) start from the same snapshot. (d) Cartoon representations of two alternative native state conformations marked by color-coded circles in (a)−(c). Sticks highlight specific residues.

unimodal. This is due to the presence of many irrelevant and weakly coupled features. As a consequence, no threshold can be defined to approximately separate the native state from unfolded conformations, *i.e.*, nearest neighbor relations become meaningless.[1,10] Upon utilizing global weights, slow features have more influence, and the distribution has several distinct peaks that can be associated with native and unfolded conformations, respectively. We emphasize that a featureless distance spectrum is a fundamental and not merely a statistical problem, *i.e.*, it is not rectifiable by increasing the overall sampling density.

Figure 5b documents that on short time scales (<10 ns) the GW measure yields higher values for the autocorrelation of the corresponding distance time series than the UW measure. This result implies that kinetic proximity can be represented more accurately by weighted distance functions. The grouping or ordering of snapshots to reveal kinetically homogeneous states is precisely what Markov models,[24] diffusion maps,[31,32,43,44] cut-based free energy profiles[23] (see Figure S1), or SAPPHIRE plots[34] try to accomplish. The latter are an efficient tool for the analysis and visualization of long MD trajectories. SAPPHIRE plots offer an intuitive illustration of the states and sequence of events encountered during the simulation (see Methods), and we have previously used SAPPHIRE plots to analyze data from MD simulations of protein folding,[34,35] the conformational dynamics of proteins,[35,45] and the binding of a peptide to a protein domain.[46] We next use the method to further evaluate the discriminatory ability of the UW, GW, and LAW measures.

Figure 6 shows SAPPHIRE plots based on all 3 measures. The time lag for the GW measure was set to τ = 2 ns, and we used Δ = 2 ns and α = 1 for the LAW measure. All profiles start from a snapshot in the native basin of Beta3S (see Supporting Methods for further details). The UW measure is unable to discriminate between kinetically similar and dissimilar snapshots, which leads to a relatively featureless profile (Figure 6a). The low height of the folding barrier at a progress index value of 4×10^5 indicates that the cutting surface does not delineate metastable states accurately. With weights, higher barriers are obtained everywhere, and several metastable states can be detected besides the native state (Figures 6b and 6c). We use a secondary structure annotation resolved by residue that is based on the DSSP algorithm[40] to confirm that the individual basins correspond to distinct conformations of the peptide. For weighted distance functions, the kinetic annotation and the sampling time reveal substructure in the native state of Beta3S, some of which is the result of the dynamics of the χ_1 angle of Trp10. This side chain samples two distinct conformations within the native state as shown in Figure 6d. Previous analyses did not capture this partitioning of the native basin because the relevant features were omitted or because their effective weight was too low.[20,34,36,47−49] We show in Figure S1 that a

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

backbone-centric RMSD distance places both of the conformations in Figure 6d in exactly the same basin.

In summary, Figure 5a illustrates why the inclusion of all features with equal weight is generally infeasible. We provide evidence for this in the context of three different unsupervised learning protocols (Figures 3, 4, 6, and S1). Our observations also point to the risks incurred by manual feature selection. Specifically for the mining of MD data, the primary risk lies in lumping kinetically separable states together as has happened for the native state of Beta3S in prior analyses. We believe that our approach of weighting the individual features according to kinetic information is a suitable compromise between these two extremes.

BPTI. We next analyzed the simulated dynamics of the 58-residue protein BPTI as reported in a very long MD trajectory containing 41250 snapshots saved every 25 ns.[37] In these data, BPTI explores several distinct, native-like states interconverting on the μs time scale. Compared to Beta3S, the data are of higher dimensionality, yet the overall variance is smaller.

To illustrate that angles decay on a wide range of time scales, Figure 7a plots the autocorrelation functions of selected dihedral angles. The time series of the slow χ_1 and ψ angles of Cys14 and Arg42, respectively, show that these angles likely report on metastable states, *i.e.*, jumps in these dihedral angles coincide with jumps in the RMSD time series (Figure S2). In contrast, no such conclusion is obtained for the time series of a fast angle, *e.g.*, the χ_1 angle of Leu6. These observations corroborate our hypothesis that slow degrees of freedom are more relevant than fast ones. In Figure 7b, we plot the global weights required for the GW measure. The data confirm that the slow dynamics are generally governed by the anchor points of the Cys14-Cys38 disulfide bond as well as their immediate surroundings and by the N-terminal helix.[37,50] Interestingly, the ω angle between Cys14 and Lys15 includes a component on the high μs time scale even though the peptide bond does not isomerize during the runs (Figure S2). A cartoon illustration of BPTI highlighting the slowest residues is given in Figure 7c.

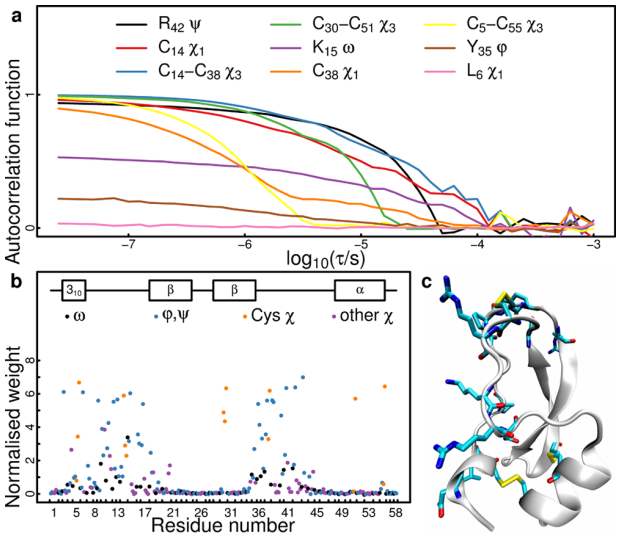


Figure 7. Autocorrelation functions and derived weights for BPTI. (a) Autocorrelation functions of selected dihedral angles. For each dihedral angle, the autocorrelation function was computed as the maximum of the autocorrelation functions of its sine and cosine values. (b) Weights $w_i = \max(R_i(1\ \mu s), 0)$ for 271 nonsymmetric dihedral angles including χ_2 and χ_3 angles of cysteines. Here, R_i is the autocorrelation function of the i^{th} dihedral angle as in (a). The weights were normalized such that their average is one and are ordered first by residue and then by type (ω, φ, ψ, χ_1, ..., χ_n) from left to right. Weights pertaining to χ_3 angles of disulfide bonds are assigned to the cysteine with lower residue number. The weights are colored according to the type of the corresponding dihedral angle. Secondary structure elements found in the crystal structure (PDB ID 5PTI)[51] are indicated on top. (c) Cartoon illustration of the crystal structure of BPTI. The residues having at least one dihedral angle with a normalized weight above 5 are shown in a stick-like representation. The illustration was rendered with VMD.[42]

Without weights (UW measure), we anticipate that a distance function based on dihedral angles is unable to reveal the conformational states of BPTI. Irrelevant features such as the χ_1 angle of Leu6 are expected to outweigh the impact of important features such as the ψ angle of Arg42 (Figure 7a). The SAPPHIRE plot[35] shown in Figure 8a confirms this prediction. The kinetic profile lacks significant barriers. With the help of the structural and sampling time annotations, 2 major and possibly 1 to 3 minor states can be identified. About 30% of the data seem to correspond to a heterogeneous ensemble. Figure 8b demonstrates that this interpretation is erroneous. The GW measure with $\tau = 1\ \mu s$ allows the kinetic and time series annotations to unmask several metastable states that are structurally distinct. The notion of a heterogeneous state with rapid interconversion is lost. The most populated basin ranges from progress index values of about 1 to 16500. The second most populated basin, found at progress index values of about 24000 to 32500, contains those snapshots most similar to the crystal structure (PDB ID 5PTI).[51] Both major basins are observed directly in NMR experiments, albeit with different weights.[37,52]

The structural annotation in Figure 8b highlights that the two major states and the conformations located between progress index values of about 21500 and 24000 all exhibit different arrangements of the Cys14−Cys38 disulfide bond. In fact, this disulfide bond has been the focal point of several studies of the native state dynamics of BPTI.[41,52,53] In particular, Xue et al. used the same MD simulation data in order to improve interpretation of data from NMR relaxation dispersion measurements.[41] They defined conformational states based on the side chain dihedral angles of Cys14 and Cys38, thus neglecting all other degrees of freedom. Comparison with their state decomposition as shown in Figure 8b demonstrates that the conformational space of BPTI is captured surprisingly well with these dihedral angles alone. The five well-defined states do not overlap significantly. This annotation also highlights the poor performance of the UW measure as seen in Figure 8a. It is of course expected that states differing in other parts of the protein are likely to be missed by the manual feature selection of Xue et al. For example, the distinct basins between progress index values of about 18500 and 21500 would be annotated as either the most populated state (red) or as unclassified (gray) by Xue et al. Similarly, the small basin at progress index values of around 33500 would be annotated as crystal-like (blue), yet it differs from the crystal structure in the orientation of the Cys30−Cys51 disulfide bond. We note that the analysis of Shaw et al.[37] also failed to separate these minor states despite selecting features that are putatively sensitive to them (separate annotation in Figure 8).

In Figure 8c we study the same data set using the LAW measure ($\Delta = 1\ \mu s$, $\alpha = 1$). The resultant picture is very similar to the one based on the GW measure. We hypothesize that the

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

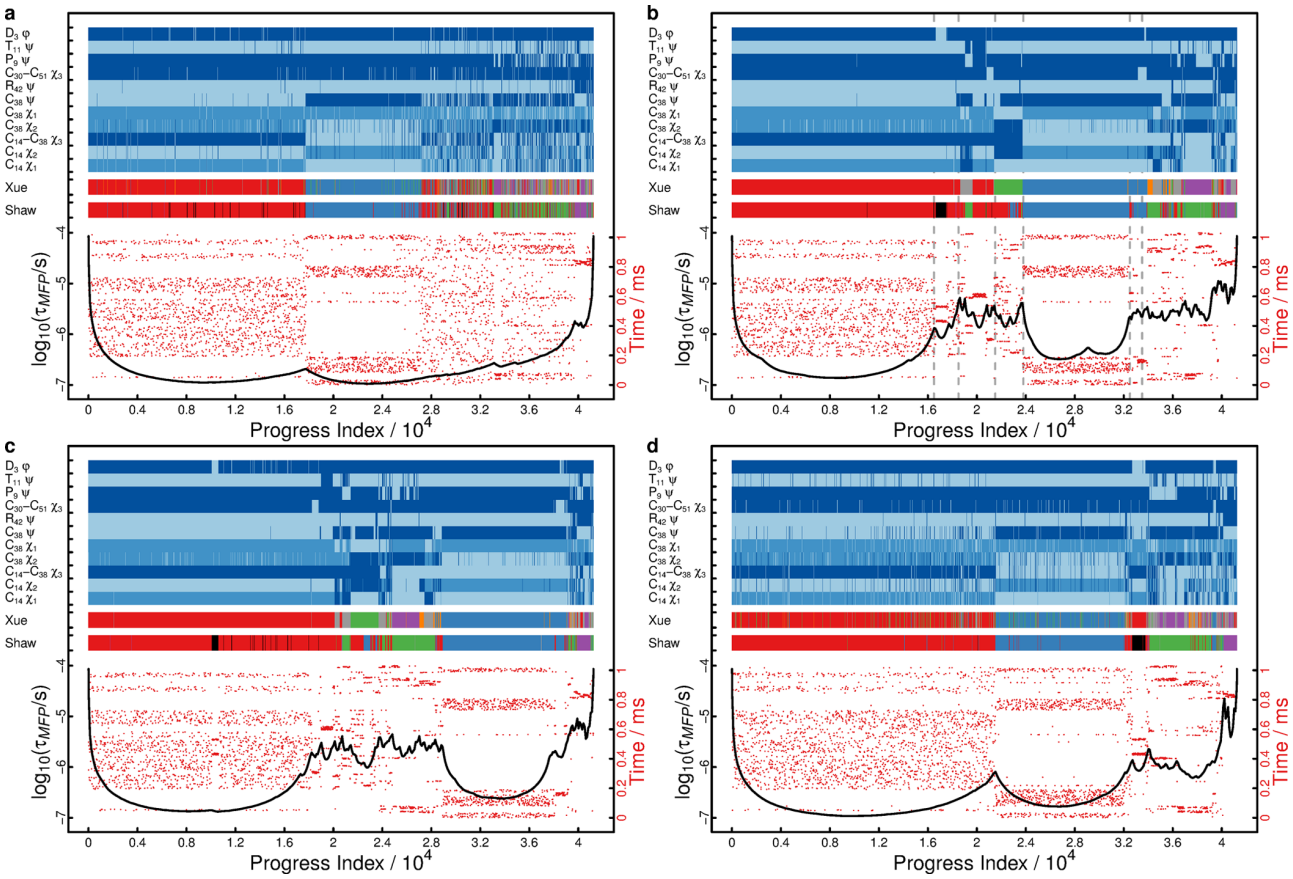


Figure 8. SAPPHIRE plots for folded BPTI. We use the same algorithm[34] as in Figure 6. (a) The (unweighted) Euclidean distance of the sine and cosine values of 271 dihedral angles (UW measure) is used to generate the progress index (x axis), which is annotated with kinetic information (black curve), sampling time (red dots), and structural information (color annotation on top). We extend this SAPPHIRE plot[35] by color-coded state assignments according to Shaw et al.[37] (red, blue, green, magenta, and black for states 0 to 5) and Xue et al.[41] (M1 - blue, M2 - orange, M3 - magenta, m_{C14} - red, m_{C38} - green, and other states - gray). The color-coded structural information uses binning of selected dihedral angles for clarity (see Methods). All annotations except the kinetic one are plotted every 4th snapshot to keep the size of the original vector image manageable. (b) The same as (a) for the GW measure with $\tau = 1$ μs. Dashed, gray lines indicate features of the plot discussed in the text. (c) The same as (a) for the LAW measure with $\Delta = 1$ μs and $\alpha = 1$. (d) The same as (a) using the RMSD of all 699 nonsymmetric atoms as the distance function.

relevance of individual features stays roughly the same throughout the conformational space sampled, which is composed of metastable states with high mutual similarity. Consequently, a locally adaptive distance function is not essential for this particular system. Figure 8d demonstrates that a SAPPHIRE plot employing the RMSD of all nonsymmetric atoms as the distance functions captures most states. However, it does not capture the m_{C38} (green) state in the model of Xue et al., a conformation for which there is experimental evidence.[41,52]

An obvious question to ask regards the dependency of our approach on the time domain parameters, τ (GW) and Δ (LAW), and the accuracy of the derived weights. It is well-known that estimators of second moments or related quantities have poor convergence properties with the numbers of samples. For time correlation measures, this is exacerbated in cases where the raw data do not sample the span of the underlying distribution recurrently. Surprisingly, Figure S3 demonstrates that the results obtained with the GW measures are largely preserved even with radically different choices for the parameter τ. This indicates that the main benefit of the GW measure for BPTI lies in reducing the influence of fast dihedral angles (compare Figures 7a and S2). Conversely, it appears to be less important how the slow dihedral angles are weighted with respect to one another. This is critical since it means that an accurate estimation of the true autocorrelation function at fixed lag time is not needed, thus preserving applicability of the method to cases where sampling is still poor (several smaller states in both Figures 6 and 8 are visited only once as indicated by the time series annotations).

We can take this point further by considering a very slow backbone dihedral angle that undergoes no significant transition for a given finite data set. Due to slow modes not actually being sampled, the autocorrelation function at large enough lag time would in all likelihood be close to zero, thereby giving a very slow degree of freedom a negligible weight. However, this seemingly misleading result is beneficial for the analysis as it reduces the weight of a feature containing no useful information (lack of variance). This example illustrates that the weights in the GW measure are data-driven, *i.e.*, they respond meaningfully to the finite samples available and need not be informed by the true distribution in a hypothetical limit. The same argument can be extended to the LAW measure using the same example. The data-driven origin of weights also implies that simulations using biased Hamiltonians, *e.g.*, umbrella sampling,[54] can be analyzed in the same way as

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

shown here. The caveat that the true dynamics are unlikely to be represented faithfully by the data does not concern the analysis *per se*. Other simulation approaches yield ensembles of short trajectories at a given condition, *e.g.*, the replica exchange method.[55] Trajectory ensembles mean that limited amounts of data are available for inferring the time correlations underlying the GW and LAW measures, which is exacerbated for large lag times (τ) and window sizes (Δ). We are currently investigating the use of these measures with small values for τ and Δ in the context of a recent trajectory ensemble sampling method.[56]

To summarize, Figure 8 provides evidence that the weighted GW and LAW measures provide a richer picture of the conformational space of BPTI than two reference approaches, *viz.*, the use of (nearly) complete sets of features with equivalent weights for either dihedral angles or coordinate RMSD. We show that both of the latter bear the risk of lumping distinct states together. We characterize these states as distinct because they are structurally and kinetically homogeneous as highlighted by the annotations in Figure 8. An accurate definition of states is required to appropriately study state-dependent processes such as the exchange of the internal water molecules of BPTI.[57]

4. DISCUSSION

We propose to weight features in the evaluation of the distance between high-dimensional vectors, *e.g.*, dihedral angles recorded along MD trajectories. Specifically, the two approaches introduced are as follows. For the GW measure (eq 2), we globally weight (*i.e.*, scale) individual features according to the autocorrelation function at fixed lag time. For the LAW measure (eq 4), we count transitions across the global mean of a feature in a time-local window. Both approaches are designed to enhance the influence of slowly varying degrees of freedom. We have provided evidence that both the GW and LAW measures improve the quality of information that can be extracted from large sets of MD snapshots. The weighted distance functions have been tested on a 9-dimensional model system and on two data sets from MD simulations in conjunction with different unsupervised learning methods. The feature weighting method has unmasked slow dynamics of side chain packing within the native state of Beta3S (Figure 6) and revealed metastable conformations of BPTI that were not resolved in previous analyses (Figure 8).

A significant advantage of our method is that it is predominantly data-driven, *i.e.*, little prior knowledge about the system is needed, and potential sources of human bias are eliminated. The weighted distance functions reduce the impact of features lacking or failing to sample slow modes. The weights do not correct for heterogeneous variances and cross-correlation effects. The use of dihedral angles may be advantageous in both regards. The GW measure is expected to define a metric space offering increased contrast between similar and dissimilar data points for data of this type. The same holds for the LAW measure with the caveat that the rigorous notion of a metric is lost (see Methods). The method is easy to implement, and the weights can be computed in linear time with respect to the number of snapshots. Evaluating the resulting distance function scales linearly with the dimensionality of the data. For the LAW measure, the major limitation is given by the saving frequency, which has to be high enough to resolve state-specific fluctuations over a time window that does not exceed lifetimes of the states of interest. This limitation is a result of using time (rather than geometric) locality to derive locally adaptive weights.

Related work on distance learning for MD data ranges from manual and application-specific feature selection to defining new functional forms[58,59] and modifying classical methods for dimensionality reduction. Among the latter, sketch-map is a version of multidimensional scaling[33] focusing on matching intermediate distances.[29] If the distance function used in the original high-dimensional space lacks contrast, such an intermediate distance separating similar from dissimilar data points cannot be defined. Locally scaled diffusion map is an extension of diffusion map[30,31,43] using a Gaussian kernel with data point-dependent local scales.[32] At present, both methods give equal weight to all the original features no matter how noisy or irrelevant they are, which might reduce their effectiveness in capturing the kinetics of the system if no additional feature selection is performed.

The so-called time structure-based independent component analysis (TICA)[60,61] provides a linear (but not orthonormal) transformation of the input data to yield components with maximal autocorrelation function at a given lag time. This method is conceptually similar to our GW measure, and we provide a direct comparison in Figure 9 for an attempted embedding in 2 dimensions. Comparison of Figure 9a with Figure 9b highlights that standard signal processing tools such as variance normalization can hurt rather than help the resolution of a projection based on principal component analysis (PCA). The GW measure is representable in PCA by scaling the input features according to factors of $\sqrt{w_i}$ (eq 2, τ =

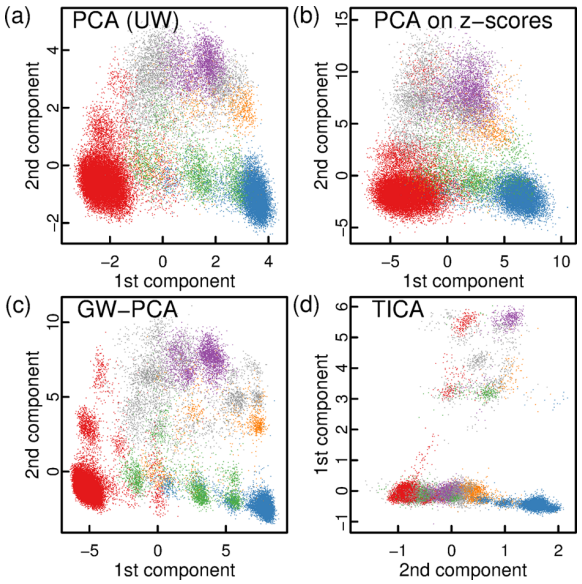


Figure 9. Two-dimensional embeddings for folded BPTI. All data points in all panels are colored according to the model introduced by Xue et al.,[41] *i.e.*, M1 - blue, M2 - orange, M3 - magenta, m_{C14} - red, m_{C38} - green, and other states - gray. (a) Projection of the data (sine and cosine values of 271 nonsymmetric dihedral angles including χ_2 and χ_3 angles of cysteines) onto the first two principal components without prior scaling of the input features. (b) The same as (a) for input features scaled to have unit variance. (c) The same as (a) for GW-PCA and a lag time of τ = 1 μs. (d) Projection of the same data onto the first two time structure-based independent components using a lag time of τ = 1 μs. Note that components are swapped to highlight similarity to other panels.

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

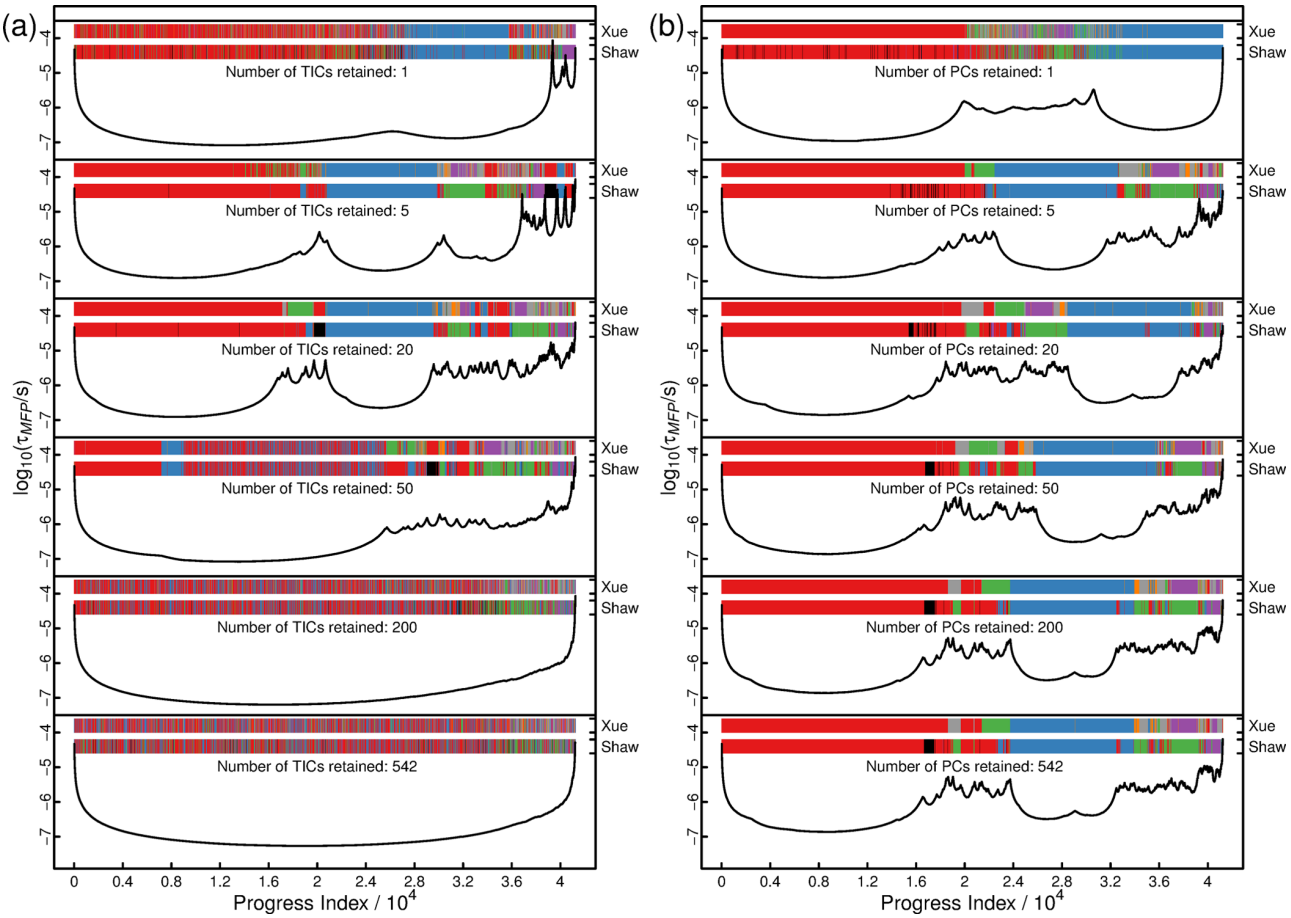


Figure 10. Comparison of TICA and GW-PCA for BPTI. Simplified SAPPHIRE plots are shown as in Figure 8 with only two annotations and the kinetic cut function plotted. (a) TICA eigenvectors and eigenvalues were computed for the raw data (Euclidean distance of the sine and cosine values of 271 dihedral angles) using a lag time of τ = 1 μs. Components were ordered by eigenvalue (value of the autocorrelation function). Data were then transformed and different numbers of those features with the largest eigenvalues were retained as indicated. (b) The same for GW-PCA. PCA was applied to data scaled by the global weights as defined for the GW measure and a lag time of τ = 1 μs. When retaining all 542 features, the progress index becomes identical to that in Figure 8b because PCA yields an orthonormal transformation.

1 μs). The resultant GW-PCA approach (Figure 9c) separates the data much better, and numerous states emerge. In contrast, the TICA approach (Figure 9d)[61] appears to overemphasize a particular slow mode leading to excellent separation of a small subpopulation but dramatic overlap of everything else. As an additional point, Figure 9 emphasizes that the usefulness of the weights is not specific to the analysis methods employed in Figures 2, 3, 6, and 8.

To test this result further, we recomputed the progress index for BPTI using both GW-PCA and TICA with a wide range of retained dimensionalities. It emerges that an appropriate choice of dimensionality is critical in TICA but not in GW-PCA (Figure 10). The best-performing TICA case retains 20 features, which correlates well with the eigenvalue spectrum (not shown). However, even this case is not obviously adding to the information provided by GW-PCA, which appears robust for most of the tested dimensionalities. Importantly, the full-dimensional GW-PCA case is equivalent to the original GW measure in Figure 8b and performs as well or better than any TICA example shown. We note that TICA fails dramatically at high dimensionality and provides even less information than the UW measure (Figure 8a).

Based on Figure 10, it is clear why recent TICA applications resort to a low-dimensional embedding of the transformed data.[38,39] A common limitation of TICA and GW-PCA is their inherent linearity although kernel-based extensions might capture nonlinear structure in the data.[62] A method similar to the LAW measure is that by Singer et al.[63,64] who propose the semimetric

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l)) = (\mathbf{x}(t_k) + \mathbf{x}(t_l))^T(\Sigma_k^{-1} + \Sigma_l^{-1})(\mathbf{x}(t_k) + \mathbf{x}(t_l))$$

where Σ_k is a local covariance matrix associated with $\mathbf{x}(t_k)$. It can be determined by running short stochastic simulations starting from $\mathbf{x}(t_k)$,[63,64] which is not feasible for the large MD data sets considered in the present study, or from the data within a short time window along the trajectory around $\mathbf{x}(t_k)$,[65] similar to what we have proposed here (eqs 3 and 4). Other approaches directly take advantage of kinetic information to determine relevant features before applying any unsupervised learning protocol. The method of McGibbon and Pande learns a distance function that tends to return low values for pairs of data points that are close in time and large values for those pairs that are far in time along the trajectory.[66] This task is formulated as a complex optimization problem depending on several parameters. When the approach was applied to MD data

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

of the protein Fip35, side chain and peptide bond dihedral angles were discarded *a priori*, suggesting that manual feature selection is still needed.

All the algorithms used here have been implemented in the free software package CAMPARI (http://campari.sourceforge.net), and the current development version is available upon request (campari.software@gmail.com). Ongoing work is focused on combining the framework of weighted distances with the RMSD metric in order to study processes involving multiple molecules such as ligand binding to receptors. However, external motion complicates the definition of weights for atoms.

In conclusion, we have developed a data-driven method for feature weighting to improve the contrast of distance functions. Our method reveals metastable states in the reversible folding of Beta3s and the native state of BPTI, which were not resolved in previous studies of the same data sets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00618.

Supporting Methods with values for auxiliary parameters and further data on Beta3S (Figure S1) and BPTI (Figures S2 and S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: a.vitalis@bioc.uzh.ch.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank David E. Shaw for sharing the trajectory data and the state annotation for BPTI (color code in Figures 8 and 10). This work was supported by a grant from the Swiss National Science Foundation to A.C. and a fellowship from the Holcim Stiftung Wissen to A.V.

■ REFERENCES

(1) Clarke, R.; Ressom, H. W.; Wang, A.; Xuan, J.; Liu, M. C.; Gehan, E. A.; Wang, Y. The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **2008**, *8*, 37−49.

(2) Allison, D. B.; Cui, X.; Page, G. P.; Sabripour, M. Microarray data analysis: From disarray to consolidation and consensus. *Nat. Rev. Genet.* **2006**, *7*, 55−65.

(3) Hilario, M.; Kalousis, A. Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings Bioinf.* **2008**, *9*, 102−118.

(4) Bhat, P. C. Multivariate analysis methods in particle physics. *Annu. Rev. Nucl. Part. Sci.* **2011**, *61*, 281−309.

(5) Beyer, K. S.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is "nearest neighbor" meaningful? In *ICDT '99, Proceedings of the 7th International Conference on Database Theory*, Jerusalem, Israel, January 10−12, 1999; Beeri, C., Buneman, P., Eds.; Springer: Berlin, 1999; pp 217−235.

(6) Hinneburg, A.; Aggarwal, C. C.; Keim, D. A. What is the nearest neighbor in high dimensional spaces? In *VLDB 2000, Proceedings of the 26th International Conference on Very Large Data Bases*, Cairo, Egypt, September 10−14, 2000; El Abbadi, A., Brodie, M. L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., Whang, K.-Y., Eds.; Morgan Kaufmann: Orlando, USA, 2000; pp 506−515.

(7) Kriegel, H. P.; Kröger, P.; Zimek, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discovery Data* **2009**, *3*, 1.

(8) Domeniconi, C.; Gunopulos, D.; Ma, S.; Yan, B.; Al-Razgan, M.; Papadopoulos, D. Locally adaptive metrics for clustering high dimensional data. *Data Min. Knowl. Discovery* **2007**, *14*, 63−97.

(9) Domeniconi, C.; Papadopoulos, D.; Gunopulos, D.; Ma, S. Subspace clustering of high dimensional data. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, Lake Buena Vista, FL, USA, April 22−24, 2004; Berry, M. W., Dayal, U., Kamath, C., Skillicorn, D. B., Eds.; SIAM: 2004; pp 517−521.

(10) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, 2009; pp 485−698.

(11) Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651−666.

(12) Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78−87.

(13) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.

(14) Jolliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002; pp 1−405.

(15) Xu, R.; Damelin, S.; Nadler, B.; Wunsch, D. C. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. *Artif. Intell. Med.* **2010**, *48*, 91−98.

(16) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064−4092.

(17) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein-folding simulations. *Nat. Phys.* **2010**, *6*, 751−758.

(18) Buchner, G. S.; Murphy, R. D.; Buchete, N. V.; Kubelka, J. Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory. *Biochim. Biophys. Acta, Proteins Proteomics* **2011**, *1814*, 1001−1020.

(19) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.

(20) Vitalis, A.; Caflisch, A. Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theory Comput.* **2012**, *8*, 1108−1120.

(21) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495−1517.

(22) Rao, F.; Caflisch, A. The protein folding network. *J. Mol. Biol.* **2004**, *342*, 299−306.

(23) Krivov, S. V.; Karplus, M. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* **2006**, *110*, 12689−12698.

(24) Noé, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.

(25) Buchete, N. V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(26) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 45−52.

(27) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−2323.

(28) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9885−9890.

(29) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

(30) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5−30.
(31) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.* **2008**, *7*, 842−864.
(32) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 124116.
(33) Cox, M. A. A.; Cox, T. F. Multidimensional scaling. In *Handbook of Data Visualization*; Chen, C.-h., Härdle, W. K., Unwin, A., Eds.; Springer: Berlin, 2008; pp 315−347.
(34) Blöchliger, N.; Vitalis, A.; Caflisch, A. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.* **2013**, *184*, 2446−2453.
(35) Blöchliger, N.; Vitalis, A.; Caflisch, A. High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.* **2014**, *4*, 6264.
(36) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.
(37) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.
(38) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(39) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.
(40) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(41) Xue, Y.; Ward, J. M.; Yuwen, T.; Podkorytov, I. S.; Skrynnikov, N. R. Microsecond time-scale conformational exchange in proteins: Using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. *J. Am. Chem. Soc.* **2012**, *134*, 2555−2562.
(42) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(43) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 113−127.
(44) Xue, Y.; Ludovice, P. J.; Grover, M. A.; Nedialkova, L. V.; Dsilva, C. J.; Kevrekidis, I. G. State reduction in molecular simulations. *Comput. Chem. Eng.* **2013**, *51*, 102−110.
(45) Huang, D.; Caflisch, A. Evolutionary conserved Tyr169 stabilizes the β2-α2 loop of the prion protein. *J. Am. Chem. Soc.* **2015**, *137*, 2948−2957.
(46) Blöchliger, N.; Xu, M.; Caflisch, A. Peptide binding to a PDZ domain by electrostatic steering via nonnative salt bridges. *Biophys. J.* **2015**, *108*, 2362−2370.
(47) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. Extracting physically intuitive reaction coordinates from transition networks of a β-sheet miniprotein. *J. Phys. Chem. B* **2010**, *114*, 6979−6989.
(48) Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A.; Dinner, A. R.; Clementi, C. Delineation of folding pathways of a β-sheet miniprotein. *J. Phys. Chem. B* **2011**, *115*, 13065−13074.
(49) Kalgin, I. V.; Caflisch, A.; Chekmarev, S. F.; Karplus, M. New insights into the folding of a β-sheet miniprotein in a reduced space of collective hydrogen bond variables: Application to a hydrodynamic analysis of the folding flow. *J. Phys. Chem. B* **2013**, *117*, 6092−6105.
(50) Long, D.; Brüschweiler, R. Atomistic kinetic model for population shift and allostery in biomolecules. *J. Am. Chem. Soc.* **2011**, *133*, 18999−19005.
(51) Wlodawer, A.; Walter, J.; Huber, R.; Sjölin, L. Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Biol.* **1984**, *180*, 301−329.
(52) Grey, M. J.; Wang, C.; Palmer, A. G., III Disulfide bond isomerization in basic pancreatic trypsin inhibitor: Multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J. Am. Chem. Soc.* **2003**, *125*, 14324−14335.
(53) Otting, G.; Liepinsh, E.; Wüthrich, K. Disulfide bond isomerization in BPTI and BPTI(G36S): An NMR study of correlated mobility in proteins. *Biochemistry* **1993**, *32*, 3571−3582.
(54) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(55) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(56) Bacci, M.; Vitalis, A.; Caflisch, A. A molecular simulation protocol to avoid sampling redundancy and discover new states. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*, 889−902.
(57) Persson, F.; Halle, B. Transient access to the protein interior: Simulation versus NMR. *J. Am. Chem. Soc.* **2013**, *135*, 8735−8748.
(58) Cossio, P.; Laio, A.; Pietrucci, F. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Phys. Chem. Chem. Phys.* **2011**, *13*, 10421−10425.
(59) Zhou, T.; Caflisch, A. Distribution of reciprocal of interatomic distances: A fast structural metric. *J. Chem. Theory Comput.* **2012**, *8*, 2930−2937.
(60) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.
(61) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134*, 065101.
(62) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.
(63) Singer, A.; Coifman, R. R. Non-linear independent component analysis with diffusion maps. *Appl. Comput. Harmon. Anal.* **2008**, *25*, 226−239.
(64) Singer, A.; Erban, R.; Kevrekidis, I. G.; Coifman, R. R. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 16090−16095.
(65) Dsilva, C. J.; Talmon, R.; Rabin, N.; Coifman, R. R.; Kevrekidis, I. G. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *J. Chem. Phys.* **2013**, *139*, 184109.
(66) McGibbon, R. T.; Pande, V. S. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900−2906.

DOI: 10.1021/acs.jctc.5b00618
J. Chem. Theory Comput. 2015, 11, 5481−5492

Supporting Information for:
Weighted distance functions improve analysis of high-dimensional data: application to molecular dynamics simulations

Nicolas Blöchliger, Amedeo Caflisch, and Andreas Vitalis*

University of Zurich
Department of Biochemistry
Winterthurerstrasse 190, CH-8057 Zurich

* To whom correspondence should be addressed: a.vitalis@bioc.uzh.ch

Supporting Methods

SAPPHIRE plot for Beta3S (Figure 6 in the main text). We represented the peptide by the sine and cosine values of 99 nonsymmetric dihedral angles. We used a stochastic, approximate algorithm[1] to generate the progress indices for the SAPPHIRE plots in Figure 6. The stochastic algorithm is scalable to large data sets because of the preorganization of the data via tree-based, hierarchical clustering.[2] The upper threshold radius and the tree height for the clustering were set to 1 and 8. The lower threshold radius was set to 0.487, 0.433, and 0.449 for the SAPPHIRE plots based on the UW (eq 1), GW (eq2), and LAW (eq 4) measures, respectively. These settings were chosen to have roughly 100000 clusters at the leaf-level. All the SAPPHIRE plots use snapshot 468441 as the starting snapshot. The number of guesses to find near neighbors[1] was set to 4000. We made use of two recent improvements to the algorithm for generating the approximate progress index (Vitalis, manuscript submitted). First, after the initial clustering of the data, we cluster the data on the three levels of finest resolution again. This improves the homogeneity in the clustering on these levels. The algorithm for generating the approximate progress index requires the computation of near neighbors for the individual snapshots, and the hierarchical clustering is used to focus the search-space. Here, we allow to enlarge this search space if the number of 4000 guesses can otherwise not be satisfied. This is controlled via the CAMPARI keyword "FMCSC_CPROGRDEPTH," which was set to 3.

SAPPHIRE plot for BPTI (Figure 8 in the main text and Figure S3). For the SAPPHIRE plots shown in Figures 8a-c, 10, and Figure S3, we represented the protein by 271 nonsymmetric dihedral angles. These include χ_2 and χ_3 angles of cysteines. We used an exact algorithm to compute the progress index for these SAPPHIRE plots. Conversely, for Figure 8d, the approximate algorithm[1] was used with the root-mean square deviation (RMSD) of the positions of 699 nonsymmetric atoms as the distance function. Parameter settings for the auxiliary clustering were taken from previous work.[3] In particular, the upper and lower threshold radii and the tree height for the clustering were set to 3.6 Å, 2.5 Å, and 4, respectively. The number of guesses to find near neighbors[1] was set to 1000, and we made use of the recent improvements to the algorithm as described above for Beta3S in the context of Figure 6.

All the SAPPHIRE plots for BPTI shown in Figures 8, 10, and Figure S3 use snapshot 20521 as the starting snapshot and the value of CAMPARI keyword FMCSC_CPROGMSTFOLD was 1. For the annotation with dihedral angles we used binning into up to three bins with boundaries chosen as follows: Cys14 χ_1 (-120°, -5°, 120°), Cys14 χ_2 (-140°, 0°, 130°), Cys14-Cys38 χ_3 (0°, 150°), Cys38 χ_2 (-155°, -105°, 120°), Cys38 χ_1 (-120°, 0°, 140°), Cys38 ψ (-120°, 80°), Arg42 ψ (-100°, 75°), Cys30-Cys51 χ_3 (0°, 150°), Pro9 ψ (-115°, 70°), Thr11 ψ (-120°, 95°), Asp3 φ (0°, 100°). These boundaries were obtained from direct inspection of the individual histograms for each angle.

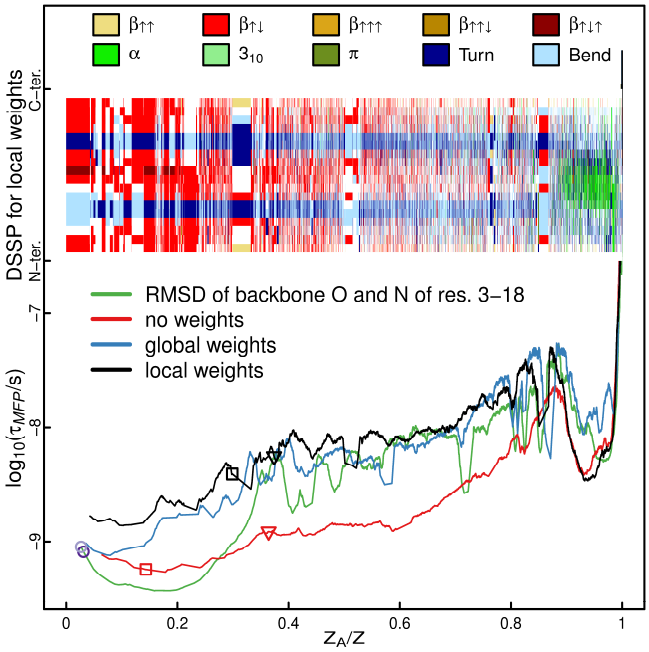


Figure S1 | Cut-based free energy profiles (cfep)[4] for Beta3S. The snapshots were clustered using a recent tree-based, hierarchical clustering algorithm[2] with four different distance functions. All the four cfeps used the largest cluster as their reference cluster. The coordinate root mean square deviation (RMSD) computed over backbone oxygen and nitrogen atoms of residues 3-18 after pairwise alignment serves as reference to illustrate manual feature selection (green curve). The clustering used coarse and fine thresholds of 10 and 1.5 Å, respectively, and yielded 161778 clusters with a tree height of 16. Note that the same parameters were used as in Figure 6A in prior work.[2] The resulting cfep displays the native state ($Z_A/Z \leq 0.38$) as a homogeneous basin without any internal structure. The second distance function we employed is the Euclidean distance of the sine and cosine values of 99 nonsymmetric dihedral angles (red curve, UW measure). The clustering used a tree height of 8, and 1 and 0.473 as coarse and fine thresholds, respectively. These settings produced 162039 clusters. The resulting cfep has a smaller folding barrier than the one based on RMSD, and the other metastable states are not as pronounced. The third and the fourth distance functions used are the GW (eq 2 in the main text, $\tau = 2$ ns) and LAW (eq 4 in the main text, $\Delta = 2$ ns, $\alpha = 1$) measures, respectively. We again chose the sine and cosine values of the same dihedral angles to represent the data (blue and black curves, respectively). Here the clustering used a tree height of 8, and a value of 1.0 as coarse threshold. The fine thresholds were set to 0.399 (GW) and 0.422 (LAW), which gave 160574 and 159754 clusters, respectively. In both cases, the folding barrier near $Z_A/Z = 0.4$ is slightly higher than for the RMSD-based profile and substructure can be found within the native basin ($Z_A/Z \leq 0.4$). The cfep based on local weights is annotated by the secondary structure[5] of the centroids of the individual clusters (legend on top). The positions of the clusters from the native basin and from the folding barrier that were extracted for Figure 3 are indicated with squares and triangles, respectively. The purple circles highlight the positions of the clusters containing the two snapshots shown in Figure 6d in the main text in the RMSD-based profile (green curve).

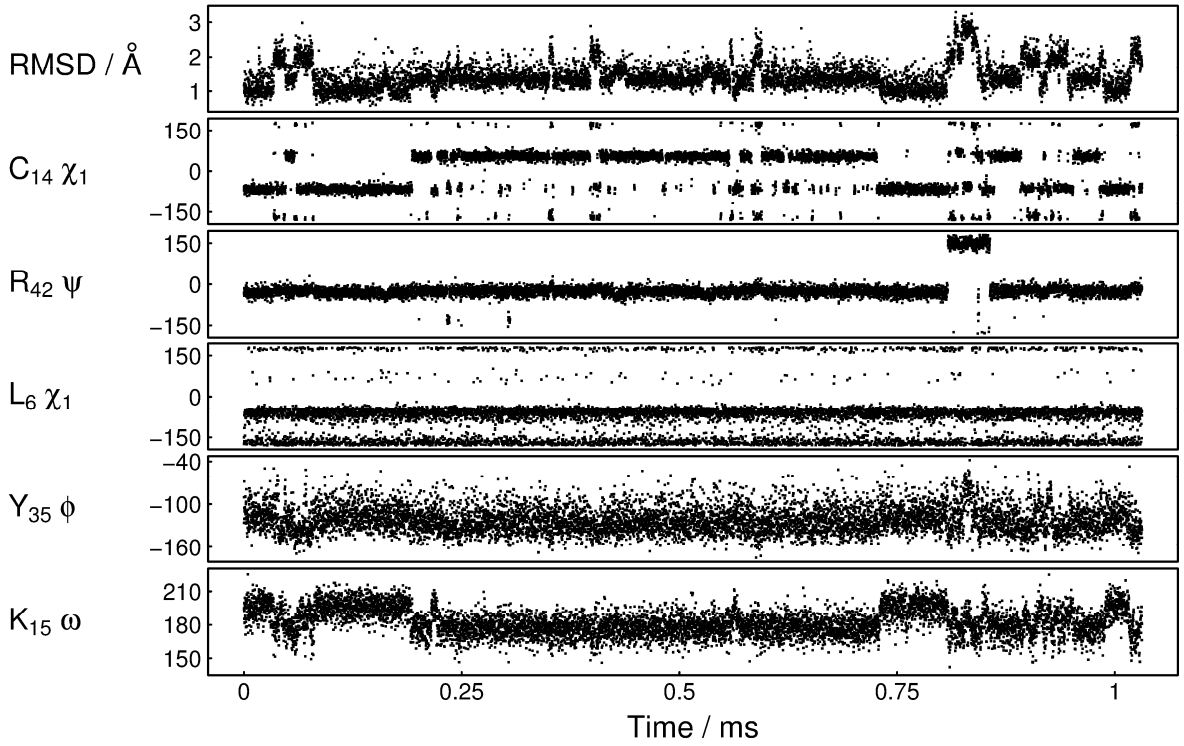


Figure S2 | Time series of the RMSD to a reference conformation and of selected dihedral angles for BPTI. For the slow dihedral angles (Cys14 χ_1, Arg42 ψ) jumps coincide with jumps in the RMSD time series. No such conclusion is obtained for fast dihedral angles (Leu6 χ_1). Other dihedral angles include a recognizable slow component but exhibit overlap among the distributions within different metastable states (Tyr35 φ and the ω angle between Cys14 and Lys15). Note that these observations are reflected in the respective autocorrelation functions (Figure 7a in the main text). Data are plotted for every 5th snapshot, *i.e.*, every 125 ns.

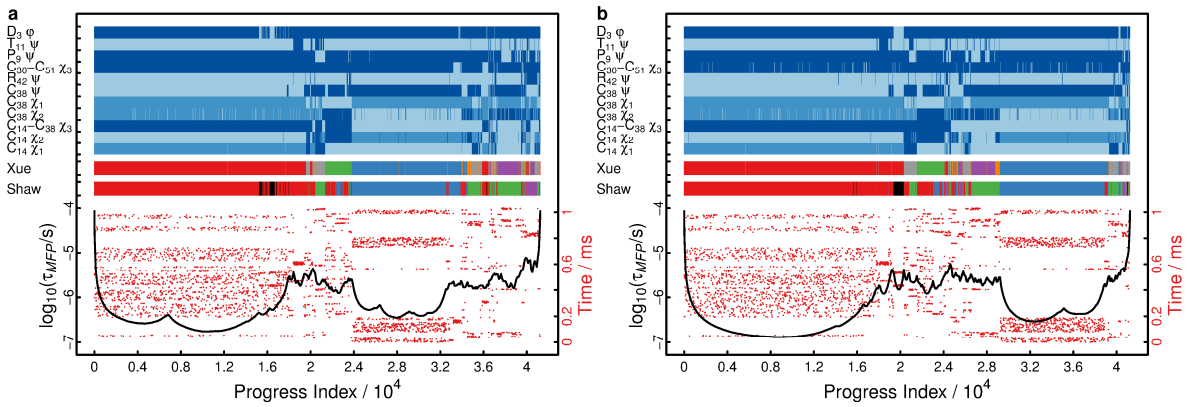


Figure S3 | Influence of the time lag τ for global weights. These SAPPHIRE plots are identical to Figure 8b in the main text with the exception that the employed lag times τ are 25 ns (**a**) and 100 µs (**b**), respectively. The saving frequency of the trajectory equals 25 ns, *i.e.*, τ = 25 ns is the lowest possible time lag. Note that most autocorrelation functions have decayed completely at τ = 100 µs (Figure 7a in the main text), which leads to very noisy estimates for the weights w_i. Nevertheless, the resulting profile highlights the most important features of the conformational landscape of BPTI. Please refer to the caption of Figure 8 in the main text for plotting details and to the Supporting Methods above for further information.

Supporting References

(1) Blöchliger, N.; Vitalis, A.; Caflisch, A., A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comp. Phys. Comm.* **2013,** *184* (11), 2446-2453.
(2) Vitalis, A.; Caflisch, A., Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theor. Comput.* **2012,** *8* (3), 1108-1120.
(3) Blöchliger, N.; Vitalis, A.; Caflisch, A., High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.* **2014,** *4*, 6264.
(4) Krivov, S. V.; Karplus, M., One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* **2006,** *110* (25), 12689-12698.
(5) Kabsch, W.; Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.

Chapter 6

Kinetic response of a photo-perturbed allosteric protein

Buchli, B., Waldauer, S.A., Walser, R., Donten, M., Pfister, R., Blöchliger, N., Steiner, S., Caflisch, A., Zerbe, O. and Hamm, P. *Proceedings of the National Academy of Sciences*, 110(29): 11725–11730, 2013

Kinetic response of a photoperturbed allosteric protein

Brigitte Buchli[a,1], Steven A. Waldauer[a,1], Reto Walser[a,1], Mateusz L. Donten[a], Rolf Pfister[a], Nicolas Blöchliger[b], Sandra Steiner[b], Amedeo Caflisch[b], Oliver Zerbe[a], and Peter Hamm[a,2]

Departments of [a]Chemistry and [b]Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Edited by Hans Frauenfelder, Los Alamos National Laboratory, Los Alamos, NM, and approved June 3, 2013 (received for review April 3, 2013)

By covalently linking an azobenzene photoswitch across the binding groove of a PDZ domain, a conformational transition, similar to the one occurring upon ligand binding to the unmodified domain, can be initiated on a picosecond timescale by a laser pulse. The protein structures have been characterized in the two photoswitch states through NMR spectroscopy and the transition between them through ultrafast IR spectroscopy and molecular dynamics simulations. The binding groove opens on a 100-ns timescale in a highly nonexponential manner, and the molecular dynamics simulations suggest that the process is governed by the rearrangement of the water network on the protein surface. We propose this rearrangement of the water network to be another possible mechanism of allostery.

Subtle conformational transitions within the folded state of highly structured proteins are often an integral aspect in their functional mechanism. These conformational transitions can occur as a result of different events, such as ligand binding, covalent modification (e.g., phosphorylation), or proteolytic cleavage. When an event or perturbation at one site in a protein changes the enzymatic activity or the binding affinity to a ligand at another distant site, this process can be described as allostery. Hemoglobin has long served as the prototypical example to study this effect, where the binding of an oxygen in one subunit modifies the affinity of binding oxygen in another subunit (1, 2). The traditional models of allostery developed by Monod et al. (3) and Koshland et al. (4) attribute allosteric effects to conformational changes by which the allosteric binding site communicates with the distant active site. There is, however, increasing evidence that allostery can be mediated also without any conformational change, relying purely on changes in internal protein dynamics (5).

PDZ domains are an important class of protein interaction modules that have been studied extensively in the context of allostery. They are found in a large variety of proteins and generally bind the C termini of their targets (6–11). As scaffolding domains, they are molecular switches that play a central role in signal transduction. For several PDZ domain proteins, allosteric interactions are an important regulatory mechanism (10, 12–15).

NMR spectroscopy has been particularly useful to elucidate the equilibrium dynamics of proteins on various timescales. The notion of allostery mediated through a change in dynamic properties was corroborated by a study of the third PDZ domain from the PSD-95/SAP90 protein (16). This protein contains an additional C-terminal α-helix (α_3), which shows no direct interaction with the peptide ligand. Removal of α_3 has a negligible effect on the structure of the PDZ core domain; however, it does lead to a large decrease in ligand binding affinity, which was shown to be entirely entropic in nature. Other studies have identified changes in (conformational) entropy of both backbone (17) and side-chain (18) dynamics in other systems to give rise to allosteric effects.

Here, we focus on the second PDZ (PDZ2) domain from human tyrosine-phosphatase 1E (hPTP1E), which has been demonstrated to possess allosteric properties (19). The PDZ domain is a small 96-residue protein with a binding groove between the α_2-helix and the β_2-strand (Fig. 1*B*). As mentioned previously, side-chain dynamics in contiguous sectors spanning the whole protein have been the proposed allosteric mechanism (19–21), but in this case, ligand binding also results in a small but significant structural change, albeit being quite small (22–24) (Fig. 1*B* and Table 1). Furthermore, a number of computational and experimental studies have addressed signal transduction pathways in the PDZ2 model system (25–34).

Allostery is the propagation of a signal between two sites of a protein. Most of the investigations so far have addressed the question of what that signal might be, e.g., a structural change vs. a change in dynamic properties. Even less is known of how such a signal propagates. Whereas NMR spectroscopy is extremely powerful in elucidating equilibrium dynamics on many timescales through relaxation experiments, its inherent time resolution is rather limited in studies of nonequilibrium processes, such as signal propagation. Transient IR spectroscopy, in contrast, provides essentially unlimited (i.e., picosecond) direct time resolution together with still significant chemical selectivity.

To make the best use of the high time resolution, it is crucial to be able to perturb the system locally and with a short laser pulse. Ideally, one would phototrigger the association or dissociation of a ligand. Here, we take an experimentally more feasible approach by covalently linking an azobenzene derivative across the binding groove, which can be switched between *cis* and *trans* isomers with a light of different wavelengths (Fig. 1*A*) (35–38). We carefully designed the system such that the structural perturbation upon isomerization of the photoswitch mimics that upon ligand binding, as is discussed in the next section. Subsequently, we use transient IR spectroscopy to investigate the nonequilbrium transition between both states and finally use nonequilibrium molecular dynamics (MD) simulations to complement the experimental results with atomistic detail.

Structural Characterization

In close analogy to ref. 39, we identified the surface-exposed amino acid pair Ser21 and Glu76 as anchor points for the photoswitch because their C_α-C_α distances in the apo and holo forms closely match those of the photoswitch length in its two configurations. Furthermore, these two residues face across the binding groove at the center. The residues were mutated to cysteines to which the photoswitch was covalently coupled (Fig. S1) (40). When searching for the anchor points, we used the NMR ensemble of structures of the PDZ2 domain [holo, 1D5G, ref. 23); apo, 3PDZ, ref. 22)], which reveals a rather large 1.5-Å change

Author contributions: A.C., O.Z., and P.H. designed research; B.B., S.A.W., R.W., M.L.D., N.B., S.S., and P.H. performed research; B.B., S.A.W., R.W., and R.P. contributed new reagents/analytic tools; B.B., S.A.W., R.W., O.Z., and P.H. analyzed data; and B.B., S.A.W., R.W., and P.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The NMR, atomic coordinates, chemical shifts, and restraints have been deposited in the Protein Data Bank, www.pdb.org (PDB ID codes 2M0Z and 2M10) and the BioMagResBank, www.bmrb.wisc.edu (accession nos. 18833 and 18834).

[1]B.B., S.A.W., and R.W. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: phamm@pci.uzh.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1306323110/-/DCSupplemental.

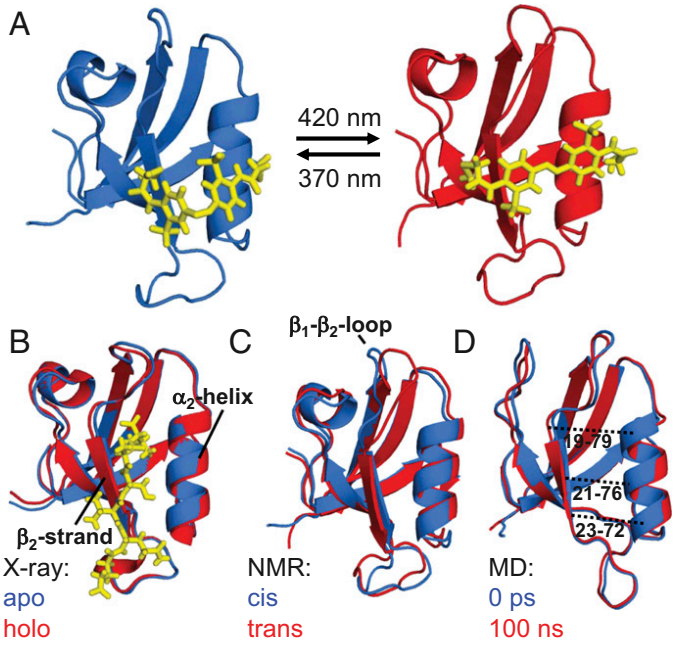


Fig. 1. (*A*) Averaged NMR structures of the photoswitchable PDZ2 domain with the photoswitch (yellow) in the *cis* (*Left*, PDB ID 2M0Z) and *trans* (*Right*, PDB ID 2M10) conformations. (*B*) Overlays of the apo (blue) and holo (red) X-ray structures [3LNX and 3LNY (24)] together with the ligand (the Ras guanine nucleotide exchange factor 2 C-terminal peptide, in yellow) in the latter case. (*C*) The NMR structures with the photoswitch *in cis* (blue) and *in trans* (red) and (*D*) the averaged MD structures with the photoswitch *in cis* (blue) and 100 ns after switching into *trans* (red). For clarity, the photoswitch is not shown in *C* and *D*. The dotted lines in *D* indicate the C_α-C_α distances shown in Fig. 4*B*.

(on average) for the Cα(21)-Cα(76) distance. Recently, X-ray crystallography (3LNX and 3LNY, ref. 24) has shown, however, that this change is about a factor of 2 smaller (i.e., 0.8 Å). The fact that the linker could nevertheless be successfully coupled and resulted in a stable, well-folded protein indicates that the possible structural perturbation is well tolerated.

The two equilibrium structures of the PDZ2 domain with the photoswitch in the *cis* and *trans* configurations were determined by NMR spectroscopy [see *SI Text* for details, Protein Data Bank (PDB) IDs 2M0Z and 2M10]. Whereas the *trans* form of the azobenzene photoswitch is predominant (~90%) after equilibration in the dark, the *cis* state was generated and maintained (>90%) by continuously illuminating the sample inside the spectrometer with a 370-nm laser coupled to a glass fiber leading into the NMR tube (for further information on the design see *SI Text*), similar to that described in previous work (41, 42). Both forms are stably folded and structurally similar to the corresponding X-ray structures (24), as can be seen in Fig. 1 *B* and *C*. More quantitatively, the rmsd of the NMR structure in *cis* compared with the X-ray structure in the apo form is 1.0 Å and that of *trans* to the holo form is 1.1 Å (considering all backbone heavy atoms of only the regions with defined secondary structure). These values include the uncertainties in the structure determination, the different environments (crystal vs. solvent), and the fact that the molecule is modified by the photoswitch. Also the adaptation of the structure upon isomerization of the photoswitch is comparable to that upon ligand binding. Table 1 lists the rmsds between the corresponding structures for all secondary structure elements and for residues of the α_2-helix and β_2-strand only, with the latter constituting the binding groove. Our construct does slightly overemphasize the conformational perturbation, but the overall agreement is quite reasonable.

Table 1. Structural comparison

rmsd	X-ray (24): apo→holo	NMR: *cis*→*trans*	MD: *cis*→*trans*
All secondary	0.34	0.92	0.46
α_2 and β_2	0.41	0.80	0.62

Structural difference of the apo vs. the holo form deduced from the X-ray structures (3LNX and 3LNY, ref. 24) or the cis vs. the trans conformer from the NMR structures and the MD simulations, respectively. The first row reports the rmsd (in angstroms) when considering all backbone atoms of regions with defined secondary structure and the second row that when considering only the α_2-helix and the β_2-strand.

Transient IR Spectroscopy

Having established the equilibrium structures of the starting and final states by NMR spectroscopy, we now turn to IR spectroscopy. Fig. 2*A* displays stationary FTIR spectra in the region of the amide I band, which is a sensitive probe of the structure of the protein backbone. All IR experiments have been performed in a fully hydrated state [50 mM borate buffer (pH 8.5) and 150 mM NaCl, lyophilized and dissolved in D_2O at 1.3 mM concentration].

A difference signal upon switching *cis* → *trans* or *trans* → *cis* is clearly visible (Fig. 2*A*, red and green) with an intensity of about 1/50th that of the absolute amide I band (Fig. 2, blue), indicating that small changes in the protein backbone do indeed occur. The *trans* → *cis* difference spectrum, induced by illuminating a dark-adapted sample at 370 nm, and the *cis* → *trans* difference spectrum, after switching off the 370-nm light and subsequent relaxation back to *trans* in ~30 min, are mirror images of each other (Fig. 2*A*, green and red lines), confirming that the molecule can be switched reversibly. In the next step, to observe the

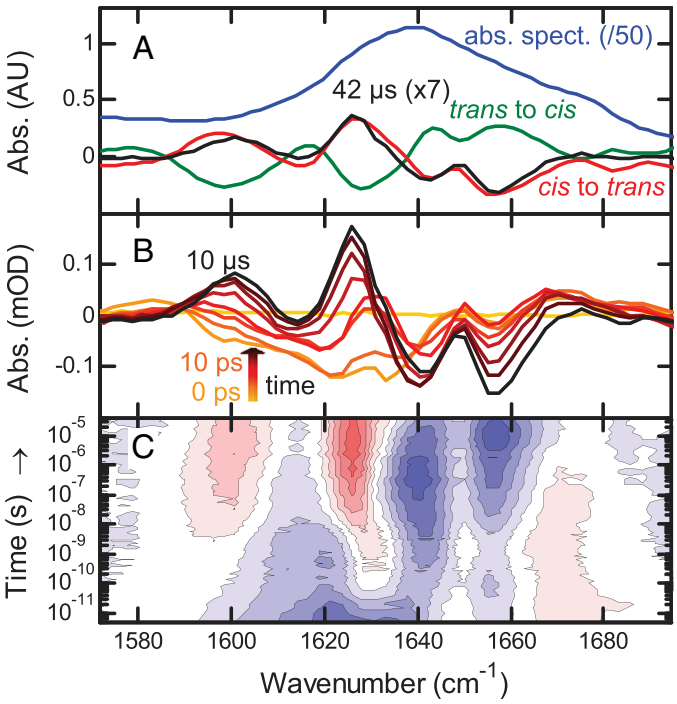


Fig. 2. (*A*) Absolute (photoswitch *in trans*, blue, downscaled by 50) and difference FTIR spectra (red and green) compared with the transient spectrum at 42 µs (black, upscaled by 7). (*B*) Transient difference spectra at −1 ns (yellow), 0 ps (light orange), and from 10 ps to 10 µs by decade (orange to black). (*C*) Contour plot of the IR response. Red indicates induced absorption, blue indicates a bleach, and contour lines are equally spaced.

transition in time-resolved experiments, we prepared all samples in the *cis* configuration by continuously illuminating them at 370 nm with an excess amount of light and then initiated the *cis* → *trans* transition with a picosecond 420-nm pulse (*Materials and Methods*). At long times (42 μs), the transient IR response in the amide I region closely resembles the shape of the steady state (Fig. 2*A*, black vs. red line), indicating that the structural transition is essentially complete after this time. Its amplitude its about one-seventh of the latter, allowing one to estimate the combined excitation probability and isomerization yield.

In reaching the final state, the data reveal a complex evolution over many orders of magnitude in time with bands that both change in intensity and shift in frequency (Fig. 2 *B* and *C*). No physically meaningful model with a limited number of discrete intermediate states could be identified to which we could globally fit the data. Similar observations have been made for downhill folding (43, 44), with different kinetic responses for different spectroscopic observables. This behavior is indicative of a continuous transition between initial and final states without any significant barriers on the pathway.

Three major phases of the overall process can nonetheless be identified. We illustrate these phases in Fig. 3 for the amide I band, choosing 1,640 cm^{-1} as probe wavelength (red circles), and a strong band at 1,491 cm^{-1} originating from the photoswitch linked to the PDZ2 domain (Fig. 3, green circles). This band is the amide II vibration of the amide unit of the linker connecting the azobenzene moiety to the protein (Fig. S1*A*). For comparison, the response of the same band of the unlinked photoswitch is shown as well (Fig. 3, blue circles; see Fig. S2, for the complete data). Phase I, clearly visible in the photoswitch bands in both the linked and unlinked form, is initiated by the absorption of a 420-nm photon and the subsequent ultrafast isomerization of the azobenzene moiety. This results in the deposition of a large amount of energy into the vibrational degrees of freedom of the molecule. The vibrational energy appears as heat and results in a broad IR signal, decaying within a few 10s of picoseconds as it quickly dissipates into the solvent (45). The heat signal happens to be zero for the amide I band at 1,640 cm^{-1}, but it is clearly visible at other probe wavelengths (Fig. 2*C*).

Subsequently, in phase II, the band of the photoswitch linked to the PDZ2 domain evolves in a highly nonexponential manner until about 100 ns (Fig. 3, green circles), significantly slower than that of the photoswitch alone (Fig. 3, blue circles).

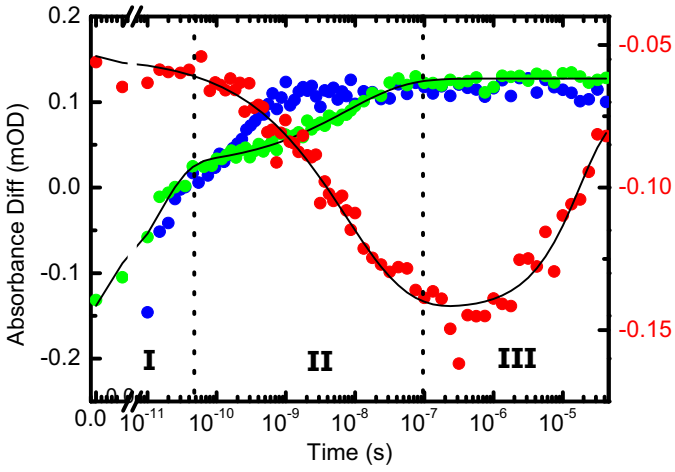


Fig. 3. Amide I response at a selected wavelength of 1,640 cm^{-1} (red, right scale) compared with that of a band at 1,491 cm^{-1} from the unlinked photoswitch (blue, left scale) and with that of the same band of the photoswitch linked to the PDZ2 domain (green, left scale). The signal of the photoswitchable PDZ2 domain (red and green) is fitted jointly to Eq. **1** (black).

Photoisomerization of the azobenzene moiety is an ultrafast picosecond process (46). The heat signal of the photoswitch linked to the protein appears on the same timescale as that of the unlinked photoswitch (Fig. S2); hence, in terms of crossing from the electronically excited back to the ground state, and thus in terms of the configuration of the central N = N bond, isomerization is equally fast in both cases. When bound to the protein, however, the photoswitch will find itself in a highly strained state after isomerization, because the protein cannot adapt instantaneously. This strain will affect also the vibrational states of the linker connecting the azobenzene moiety to the protein, i.e., the 1,491-cm^{-1} band. As the protein relaxes, the strain on the photoswitch is slowly released. Hence, the phase II signal of the photoswitch (Fig. 3, green circles) can be thought of as an indirect measure of the perturbation of the binding groove, as the binding groove is cross-linked by the photoswitch. Similar conclusions have been drawn for the electronic response of a similar azo-photoswitch in a smaller peptide (36). The perturbation of the binding groove is also reflected in the amide I band of the protein (Fig. 3, red circles).

Finally, in phase III, the photoswitch signal remains constant, because the binding groove has fully adapted to the perturbation. The amide I signal nevertheless continues to evolve in time. We assume that this signal is related to a slight rearrangement of the protein structure in a region different from the binding groove, for instance of some of the turn regions that are known to be quite flexible. This interpretation is corroborated by signal broadening of NMR resonances from the β_1-β_2 loop.

The two time traces from the photoswitchable PDZ2 domain at 1,640 cm^{-1} and at 1,491 cm^{-1} can be fitted jointly to a function (Fig. 3, black lines)

$$q(t) = a_0 + a_1 e^{-(t/\tau_1)} + a_2 e^{-(t/\tau_2)^{\beta}} + a_3 e^{-(t/\tau_3)}, \qquad [1]$$

which is composed of a fast exponential contribution for the heat signal decay in phase I (τ_1 = 15 ps), a stretched exponential contribution for the binding groove perturbation in phase II (τ_2 = 7 ns, β = 0.49), and another exponential contribution (τ_3 = 20 μs) for the final relaxation in the amide I band in phase III. In that fit, the time constants and the stretching factor were forced to be the same for both time traces, but the amplitudes were allowed to differ (Table S1). With a stretching factor of β = 0.49, the nonexponential time dependence of phase II is quite pronounced.

Nonequilibrium Molecular Dynamics

To facilitate the understanding of the transition on an atomistic level, we used nonequilibrium MD simulations in explicit water. Starting from an equilibrated *cis* ensemble (with an rmsd to the corresponding NMR structure of 1.4 Å), we launched very many nonequilibrium trajectories by instantaneously switching the potential energy function of the central N = N bond of the photoswitch from one that is stable in *cis* to one that is stable in *trans* (*Materials and Methods*) (36, 38, 47). As these simulations are limited to a maximum simulation time of 100 ns, we focus on phase II in the following discussion, i.e., the perturbation of the binding groove. The overall fold does not change during the first 100 ns after photoswitching (Fig. 1*D*), but the protein backbone is deformed slightly, as expressed by the rmsd that increases to 0.46 Å after 100 ns (Table 1 and Fig. 4*A*). The rmsd is larger when considering the α_2-helix and the β_2-strand only (0.62 Å), emphasizing that most of the structural changes occur at the binding groove on this timescale. As a function of time, the rmsd jumps relatively rapidly within the first 1 ps. It then continues to grow in a highly nonexponential manner, covering all orders of magnitude in time considered in this simulation, similar to the experimental observation (Fig. 3, green circles), and in fact it is

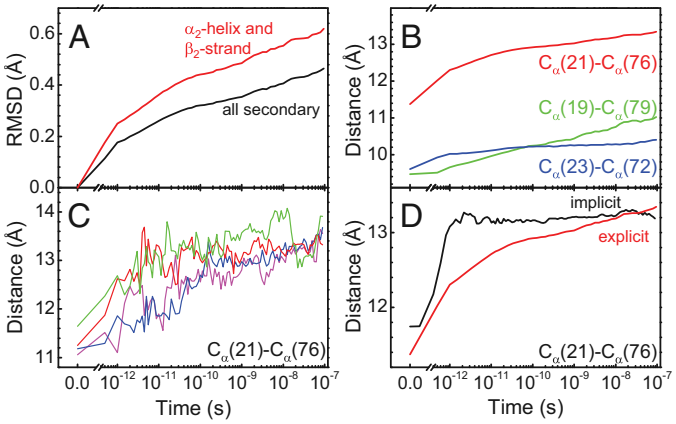


Fig. 4. MD results. (*A*) Time evolution of the rmsd, averaged over an ensemble of nonequilibrium trajectories. The rmsd is relative to the averaged starting structure. (*B*) The same for the C_α-C_α distances across the binding groove, indicated as dotted lines in Fig. 1*D*. (*C*) $C_\alpha(21)$-$C_\alpha(76)$ distance from four typical individual trajectories. *A*–*C* were all deduced from explicit water simulations. (*D*) Comparison of the $C_\alpha(21)$-$C_\alpha(76)$ distance in simulations performed in explicit (red) vs. implicit (black) water.

not finished after 100 ns. We therefore did not attempt to fit the MD data, because a stretched exponential fit becomes robust only if data exist for times long enough so that the signal levels off. No quantitative agreement is expected for the time dependence because, for example, the self-diffusion coefficient of TIP3P (three-site transferrable intermolecular potential) water used in the simulation is more than a factor of 2 higher than the experimental value (48), and water plays an important part in determining the response of the protein (see discussion below). Nevertheless, qualitatively speaking, the response is similar to phase II in Fig. 3 (green circles) [note that phase I is not directly related to any structural process, but rather to a heat signal, which is based on the anharmonicity of the molecule's potential (45) and as such is beyond the MD model]. Furthermore, the amount of structural change obtained from the MD simulation agrees reasonably well with what we find for the NMR structures (keeping in mind that the transition is not quite complete after 100 ns, Table 1).

As a more direct measure of the structural change of the binding groove, we also show in Fig. 4*B* the time evolution of various C_α-C_α distances across the binding groove. The $C_\alpha(21)$-$C_\alpha(76)$ distance, corresponding to the sites that are directly linked by the photoswitch, is perturbed the most and jumps very quickly within the first 1 ps by a significant amount of ∼1 Å. We attribute this initial jump to the direct impact of the isomerization of the photoswitch on the protein backbone. Thereafter the $C_\alpha(21)$-$C_\alpha(76)$ distance again increases steadily in a highly nonexponential manner. The neighboring pairs [Fig. 4*B*, $C_\alpha(19)$-$C_\alpha(79)$ in green and $C_\alpha(23)$-$C_\alpha(72)$ in blue] experience much less of the initial jump in change of distance, because they are not directly affected by the conformational change of the photoswitch, but the nonexponential response at later times is similar.

Nonexponential protein dynamics, modeled either as stretched exponentials or power laws, have been discussed extensively, for instance in the context of ligand (CO) dissociation and rebinding in hemoglobin or myoglobin (49–51) or the fluctuations of the pairwise distance between two sites in a protein (52–54). Two limiting scenarios of nonexponential relaxation kinetics can be distinguished (50): The parallel process is characterized by a distribution of exponential decay processes, originating from a distribution of barrier heights in an inhomogeneous ensemble of proteins. This mechanism applies in the limit when the timescale of the relaxation process is fast compared with the time the protein requires to sample its complete conformational space. In that case, individual single-molecule trajectories would still be a two-state system with either short or long C_α-C_α distances that are separated by one dominant barrier, so that one would observe essentially sudden jumps between these two states with a nonexponential distribution of jump times. Fig. 4*C* shows that we are in the opposite limit. That is, individual single-molecule trajectories essentially follow the averaged one (apart from statistical noise), without big jumps. In accordance with the conclusion already drawn from the experimental results, this observation implies that the transition is continuous without having to surmount any dominant barrier. Such a nonexponential response occurs when protein relaxation is the result of many small events, like defect diffusion that commonly leads to subdiffusive behavior (55).

Fig. 5 indicates what these many small events might be. Shown is a time series of the change of water density at the surface of the protein averaged over all nonequilibrium trajectories. As the photoswitch is isomerizing very quickly in the simulation (<1 ps), the water density changes immediately in the vicinity of the photoswitch (which is located on the right side in the structures in Fig. 5). As time proceeds, this perturbation of the water network travels around the protein and it in fact takes 100 ns until it reaches the back side. This relatively slow propagation of the perturbation of the water network needs to be put in contrast to the residence time of a given single water molecule at the protein surface, which was calculated to be 10–100 ps in the binding groove and 10–30 ps on the outside surface, respectively. Hence, water molecules need to exchange many times before a new equilibrium of the water network around the protein is established.

To gain further insights into the role of water during the structural transition, we repeated the switching simulations with implicit water (Fig. 4*D*, black line), which approximates water as a continuum of the correct dielectric constant, but without the internal degrees of freedom that would provide friction (no additional friction term was added in the Newtonian dynamics used in these simulations). Both states remain stably folded in the implicit water model, but the kinetics of the transition between them are very different from those of the explicit water model. That is, the average $C_\alpha(21)$-$C_\alpha(76)$ distance completes a 1.4-Å jump within ∼1 ps in a ballistic fashion (a minor overshoot is observed) and then stays essentially constant out to 100 ns. Clearly, the implicit water simulation is artificial and not aimed to reveal results that are comparable to reality, but is used to reveal the consequences of the direct impact of water molecules on protein dynamics through the comparison with the explicit water MD simulation. This numerical experiment shows that on the length scale of this conformational transition, the intramolecular potential of the protein alone is not rugged at all and provides no internal friction. In other words, the process is entirely "slaved" by water (51, 56, 57), and water is an integral part of the observed protein dynamics.

Conclusion

In conclusion, we have shown in a closely linked experimental-simulation study that the perturbation of the binding groove of the PDZ2 domain evolves as a continuous subdiffusive process on a 100-ns timescale. The MD simulation (Fig. 4) can essentially quantitatively reproduce the experimentally observed kinetics (Fig. 3, green circles) both in terms of its timescale and in terms of its nonexponential character. This excellent agreement validates the MD simulation, from which atomistic details may then be extracted. We find that the ruggedness of the free energy landscape that governs the dynamics of the binding groove originates entirely from water, whereas the intramolecular potential of the protein is smooth for this small conformational transition. A similar conclusion was drawn for a significantly larger conformational change between a folding intermediate

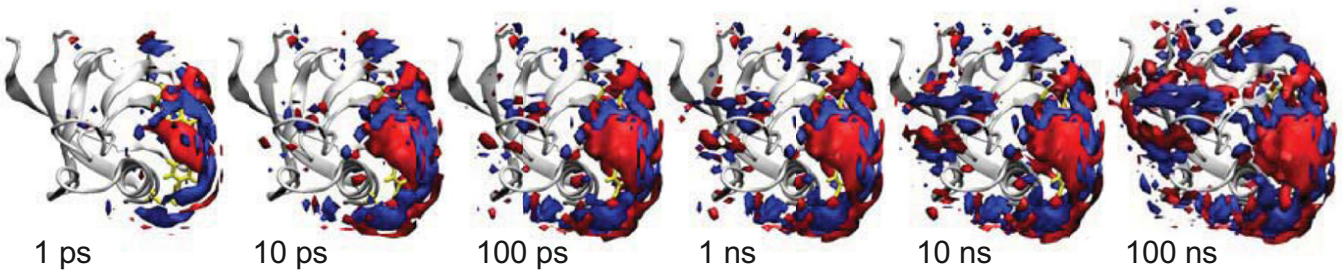


Fig. 5. Change of water density as a function of simulation time, compared with that just before switching. Red depicts increased density and blue decreased density. The contour surfaces correspond to changes of ±0.01 water/Å^3 (for comparison, the bulk water density is ∼0.033 water/Å^3). The protein is shown as a gray ribbon and the photoswitch (visible only in part) is shown in yellow. See also Movie S1.

and the native state of a four-helix bundle protein (58). We find that in our system the overall protein response is dominated by the rearrangement of the water network on the protein surface. Interestingly, the perturbation of the water network propagates around the protein within the 100 ns (Fig. 5).

We propose this to be another possible mechanism of allostery, which addresses the question of how the ligand binding site communicates with remote parts of the protein. Although the photoswitch is a rather crude mimic of ligand binding, the peptide ligand will still introduce new partial charges in the binding pocket that will rearrange the water network in its vicinity and eventually also at larger distances. This mechanism would work without any significant structural change of the protein and might even unify the seemingly competing points of view, which explain allostery either as a structural or as a dynamical effect. That is, to the extent that the dynamics of a protein are slaved by water, a change in water structure can affect the dynamics of the protein. Independent from that, phase III in Fig. 3 also hints toward a conformational change of the protein backbone in a region different from the binding groove that also could be responsible for allosteric signaling in a more traditional sense (3, 4).

Materials and Methods

Protein Preparation. The PDZ2 domain (S21C E76C) was expressed from *Escherichia coli*, using standard methods. The photoswitch 3,3′-bis(sulfonato)-4,4′- bis(chloroacetamido)azobenzene (BSBCA) was covalently linked to the two cysteines (40); see *SI Text* and Fig. S3 for details. We learned from mass spectrometry that the protein reacts photochemically with oxygen as well as with the initially used Tris buffer under the influence of the 420-nm laser pulses in the pump-probe experiment, where both presumably bind to the thioether groups of the cysteines linked to the photoswitch (Fig. S4). The experiments were therefore performed in 50 mM borate buffer (pH 8.5) and 150 mM NaCl, lyophilized and dissolved in D_2O, and care was taken to maintain the sample oxygen-free.

Time-Resolved IR Spectroscopy. Two synchronized 1-kHz Ti:sapphire oscillator/regenerative amplifier femtosecond laser systems (Spectra Physics) were used for pump-probe measurements (59). The jitter between both lasers (which effectively determines the time resolution of the experiment) was ∼10 ps and the delay could be adjusted up to 42 μs. The frequency-doubled pulses (420 nm, 3 μJ per pulse focused onto an ∼200-μm beam diameter in the sample and stretched to ∼1 ps to reduce sample deposition on the cuvette windows) of one laser system were used to excite the photoswitch. The IR probe pulses were obtained by sending the output of the second laser system through an optical parametric amplifier (100 fs, center wavenumber 1,635 cm^{-1} or 1,443 cm^{-1}). The sample was circulated in a closed-cycle flow-cell system consisting of a reservoir followed by a CaF_2 sample cell with 50 μm optical path length. The reservoir was continuously illuminated with a 150-mW, 370-nm continuous wave diode laser (CrystaLaser) so that all protein flowing into the sample cell was in the *cis* configuration.

NMR. NMR spectra of PDZ2 with the photoswitch in *trans* were recorded in the dark after equilibration. For all measurements with the photoswitch in the *cis* configuration, the sample was continuously illuminated with the 370-nm cw laser specified above (Fig. S5) coupled into the NMR spectrometer through a custom-fabricated fiber terminated with an extended cylindrical diffuser (Molex). Spectrum assignment was achieved with a standard set of triple-resonance experiments. Structure calculation was performed from NOE distance restraints from ^{15}N- and ^{13}C-resolved NOESY spectra with 75-ms mixing times. NOE data were complemented by amide proton residual dipolar couplings (NH-RDCs) measured in Pf1 bacteriophage and *n*-dodecyl-penta(ethylene glycol)/*n*-hexanol liquid crystalline medium. The 20 conformers with lowest energy of both structures showed no NOE violations bigger than 0.5 Å and showed good Ramachandran plot statistics with only 0.3% of the residues in disallowed regions (for more detailed description of experiments, parameters, and structural statistics see *SI Text* and Tables S2 and S3).

Computational Methods. MD simulations were performed with the Gromacs program package (60) and the Gromacs implementation of the Charmm27 force field (61, 62). The photoswitch was parameterized as in ref. 36. Details of the simulation protocol are given in *SI Text*.

ACKNOWLEDGMENTS. We thank Ben Schuler and his group, in particular Hagen Hofmann, for tremendous help with the protein chemistry; and Andrew Woolley and Gerhard Stock for important discussions; and the Functional Genomics Center Zurich, especially Serge Chesnov, for help with mass spectrometry. This work has primarily been supported by an European Research Council (ERC) Advanced Investigator Grant (DYNALLO) and in part by the Swiss National Science Foundation through the National Center of Competence and Research (NCCR) MUST.

1. Henry ER, Jones CM, Hofrichter J, Eaton WA (1997) Can a two-state MWC allosteric model explain hemoglobin kinetics? *Biochemistry* 36(21):6511–6528.
2. Eaton WA, Henry ER, Hofrichter J, Mozzarelli A (1999) Is cooperative oxygen binding by hemoglobin really understood? *Nat Struct Biol* 6(4):351–358.
3. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions - a plausible model. *J Mol Biol* 12:88–118.
4. Koshland DE, Jr., Némethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 5(1): 365–385.
5. Cooper A, Dryden DTF (1984) Allostery without conformational change. A plausible model. *Eur Biophys J* 11(2):103–109.
6. Songyang Z, et al. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* 275(5296):73–77.
7. Daniels DL, Cohen AR, Anderson JM, Brünger AT (1998) Crystal structure of the hCASK PDZ domain reveals the structural basis of class II PDZ domain target recognition. *Nat Struct Biol* 5(4):317–325.
8. Harris BZ, Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. *J Cell Sci* 114(Pt 18):3219–3231.
9. Jemth P, Gianni S (2007) PDZ domains: Folding and binding. *Biochemistry* 46(30): 8701–8708.
10. Lee H-J, Zheng JJ (2010) PDZ domains and their binding partners: Structure, specificity, and modification. *Cell Commun Signal* 8:8.
11. van Ham M, Hendriks W (2003) PDZ domains-glue and guide. *Mol Biol Rep* 30(2):69–82.
12. van den Berk LCJ, et al. (2007) An allosteric intramolecular PDZ-PDZ interaction modulates PTP-BL PDZ2 binding specificity. *Biochemistry* 46(47):13629–13637.
13. Li J, Callaway DJE, Bu Z (2009) Ezrin induces long-range interdomain allostery in the scaffolding protein NHERF1. *J Mol Biol* 392(1):166–180.
14. Wilken C, Kitzing K, Kurzbauer R, Ehrmann M, Clausen T (2004) Crystal structure of the DegS stress sensor: How a PDZ domain recognizes misfolded protein and activates a protease. *Cell* 117(4):483–494.
15. Whitney DS, Peterson FC, Volkman BF (2011) A conformational switch in the CRIB-PDZ module of Par-6. *Structure* 19(11):1711–1722.

16. Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL (2009) Hidden dynamic allostery in a PDZ domain. *Proc Natl Acad Sci USA* 106(43):18249–18254.
17. Popovych N, Sun S, Ebright RH, Kalodimos CG (2006) Dynamically driven protein allostery. *Nat Struct Mol Biol* 13(9):831–838.
18. Frederick KK, Marlow MS, Valentine KG, Wand AJ (2007) Conformational entropy in molecular recognition by proteins. *Nature* 448(7151):325–329.
19. Fuentes EJ, Der CJ, Lee AL (2004) Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J Mol Biol* 335(4):1105–1115.
20. Fuentes EJ, Gilmore SA, Mauldin RV, Lee AL (2006) Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. *J Mol Biol* 364(3):337–351.
21. Gianni S, et al. (2006) Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure* 14(12):1801–1809.
22. Kozlov G, Gehring K, Ekiel I (2000) Solution structure of the PDZ2 domain from human phosphatase hPTP1E and its interactions with C-terminal peptides from the Fas receptor. *Biochemistry* 39(10):2572–2580.
23. Kozlov G, Banville D, Gehring K, Ekiel I (2002) Solution structure of the PDZ2 domain from cytosolic human phosphatase hPTP1E complexed with a peptide reveals contribution of the beta2-beta3 loop to PDZ domain-ligand interactions. *J Mol Biol* 320(4):813–820.
24. Zhang J, et al. (2010) Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of protein tyrosine phosphatase 1*E*. *Biochemistry* 49(43):9280–9291.
25. Ota N, Agard DA (2005) Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J Mol Biol* 351(2):345–354.
26. De Los Rios P, et al. (2005) Functional dynamics of PDZ binding domains: A normal-mode analysis. *Biophys J* 89(1):14–21.
27. Sharp K, Skinner JJ (2006) Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Proteins* 65(2):347–361.
28. Dhulesia A, Gsponer J, Vendruscolo M (2008) Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein. *J Am Chem Soc* 130(28):8931–8939.
29. Kong Y, Karplus M (2009) Signaling pathways of PDZ2 domain: A molecular dynamics interaction correlation analysis. *Proteins* 74(1):145–154.
30. Gerek ZN, Ozkan SB (2011) Change in allosteric network affects binding affinities of PDZ domains: Analysis through perturbation response scanning. *PLoS Comput Biol* 7(10):e1002154.
31. Cilia E, Vuister GW, Lenaerts T (2012) Accurate prediction of the dynamical changes within the second PDZ domain of PTP1e. *PLoS Comput Biol* 8(11):e1002794.
32. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295–299.
33. Süel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10(1):59–69.
34. Chi CN, et al. (2008) Reassessing a sparse energetic network within a single protein domain. *Proc Natl Acad Sci USA* 105(12):4679–4684.
35. Kumita JR, Smart OS, Woolley GA (2000) Photo-control of helix content in a short peptide. *Proc Natl Acad Sci USA* 97(8):3803–3808.
36. Spörlein S, et al. (2002) Ultrafast spectroscopy reveals subnanosecond peptide conformational dynamics and validates molecular dynamics simulation. *Proc Natl Acad Sci USA* 99(12):7998–8002.
37. Rehm S, Lenz MO, Mensch S, Schwalbe H, Wachtveitl J (2005) Ultrafast spectroscopy of a photoswitchable 30-amino acid de novo synthesized peptide. *Chem Phys* 323(1): 28–35.
38. Ihalainen JA, et al. (2008) α-Helix folding in the presence of structural constraints. *Proc Natl Acad Sci USA* 105(28):9588–9593.
39. Zhang F, et al. (2009) Structure-based approach to the photocontrol of protein folding. *J Am Chem Soc* 131(6):2283–2289.
40. Burns DC, Zhang F, Woolley GA (2007) Synthesis of 3,3′-bis(sulfonato)-4,4′-bis(chloroacetamido)azobenzene and cysteine cross-linking for photo-control of protein conformation and activity. *Nat Protoc* 2(2):251–258.
41. Rubinstenn G, et al. (1998) Structural and dynamic changes of photoactive yellow protein during its photocycle in solution. *Nat Struct Biol* 5(7):568–570.
42. Kühn T, Schwalbe H (2000) Monitoring the kinetics of ion-dependent protein folding by time-resolved nmr spectroscopy at atomic resolution. *J Am Chem Soc* 122:6169–6174.
43. Garcia-Mira MM, Sadqi M, Fischer N, Sanchez-Ruiz JM, Muñoz V (2002) Experimental identification of downhill protein folding. *Science* 298(5601):2191–2195.
44. Ma H, Gruebele M (2005) Kinetics are probe-dependent during downhill folding of an engineered lambda6-85 protein. *Proc Natl Acad Sci USA* 102(7):2283–2287.
45. Hamm P, Ohline SM, Zinth W (1997) Vibrational cooling after ultrafast photoisomerization of azobenzene measured by femtosecond infrared spectroscopy. *J Chem Phys* 106:519–529.
46. Nägele T, Hoche R, Zinth W, Wachtveitl J (1997) Femtosecond photoisomerization of cis-azobenzene. *Chem Phys Lett* 272:489–495.
47. Nguyen PH, Stock G (2006) Nonequilibrium molecular dynamics simulation of a photoswitchable peptide. *Chem Phys* 323:36–44.
48. Mahoney MW, Jorgensen WL (2001) Diffusion constant of the tip5p model of liquid water. *J Chem Phys* 114:363–366.
49. Lim M, Jackson TA, Anfinrud PA (1993) Nonexponential protein relaxation: Dynamics of conformational change in myoglobin. *Proc Natl Acad Sci USA* 90(12):5801–5804.
50. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598–1603.
51. Frauenfelder H, et al. (2009) A unified model of protein dynamics. *Proc Natl Acad Sci USA* 106(13):5129–5134.
52. Volk M, et al. (1997) Peptide conformational dynamics and vibrational Stark effects following photoinitiated disulfide cleavage. *J Phys Chem B* 101:8607–8616.
53. Yang H, et al. (2003) Protein conformational dynamics probed by single-molecule electron transfer. *Science* 302(5643):262–266.
54. Milanesi L, et al. (2012) Measurement of energy landscape roughness of folded and unfolded proteins. *Proc Natl Acad Sci USA* 109(48):19563–19568.
55. Jäckle J (1986) Models of the glass transition. *Rep Prog Phys* 49:171–231.
56. Fenimore PW, Frauenfelder H, McMahon BH, Parak FG (2002) Slaving: Solvent fluctuations dominate protein dynamics and functions. *Proc Natl Acad Sci USA* 99(25): 16047–16051.
57. Vitkup D, Ringe D, Petsko GA, Karplus M (2000) Solvent mobility and the protein 'glass' transition. *Nat Struct Biol* 7(1):34–38.
58. Sekhar A, Vallurupalli P, Kay LE (2012) Folding of the four-helix bundle FF domain from a compact on-pathway intermediate state is governed predominantly by water motion. *Proc Natl Acad Sci USA* 109(47):19268–19273.
59. Bredenbeck J, Helbing J, Hamm P (2004) Continuous scanning from picoseconds to microseconds in time resolved linear and nonlinear spectroscopy. *Rev Sci Instrum* 75(11):4462.
60. Van Der Spoel D, et al. (2005) GROMACS: Fast, flexible, and free. *J Comput Chem* 26(16):1701–1718.
61. Mackerell AD, Jr., Feig M, Brooks CL, 3rd (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11):1400–1415.
62. Mackerell AD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.

Supporting Information

Buchli et al. 10.1073/pnas.1306323110

SI Text

IR-Data

The amide I band was used to observe changes in the protein backbone structure. Additionally, two bands originating from the photoswitch (see Fig. S1*A*), present in the absolute FTIR spectrum in Fig. S2*A* (blue), are sensitive to the photoswitch conformational state. The band at around 1,390 cm^{-1} is a ring mode of the azobenzene and at around 1,490 cm^{-1} is the amide II band of the two photoswitch amide groups (Fig. S1*A*). In the absolute spectra of the photoswitch bound to the protein (Fig. S2*B*, blue), these two modes are hidden behind the stronger amide II band of the protein backbone, but they still are clearly visible in the difference spectra (Fig. S2*B*, red and green), where they appear in a similar way to that for the photoswitch alone (Fig. S2*A*, red and green).

Fig. 3 (blue) of the main text shows a time trace of the photoswitch amide II band. Shortly following the heat signal decay within the first few picoseconds, the spectrum quickly relaxes to closely resemble the steady-state spectrum (Fig. 3 and Fig. S2*A*). When the photoswitch is bound to the second PDZ (PDZ2), both bands overlap with the much broader amide II band of the protein backbone (centered around 1,450 cm^{-1}; Fig. S 1*B*, blue); however, the bands originating from the photoswitch are clearly discernible in the difference spectrum. The kinetics are significantly decelerated by the counterstrain the protein imposes onto the photoswitch.

Table S1 summarizes all parameters obtained from the joint fit of the data in Fig. 3.

SI Materials and Methods

Cloning. The PDZ2 gene, containing the two mutations S21C and E76C, was synthesized and cloned into a pET30a(+)vector (EZbiolab). A human rhinovirus 3C (HRV 3C) cleavage site was added between the N-terminal hexahistidine tag and the protein gene by site-directed mutagenesis (QuikChange; Stratagene). To facilitate detection and quantification of the protein vs. the photoswitchable linker, a Trp was inserted between the His tag and the protease cleavage site (Fig. S2*B*).

Protein Expression and Purification. For IR spectroscopy measurements, the PDZ2 domain was expressed in *Escherichia coli* BL21(DE3) in Luria–Bertani medium. For NMR measurements, M9 minimal medium was supplemented with ^{13}C-glucose and/or ^{15}N-NH_4Cl. Fractional isotope labeling as described by Neri et al. (2) was used for obtaining stereospecific assignments for the Val and Leu methyl groups. The protein was purified from inclusion bodies with a HisPrep column (GE Healthcare Life Sciences) in 20 mM Tris·HCl, 6 M GdmHCl, and 10 mM imidazole, pH 8, and eluted with an imidazole gradient. Residual nickel was removed from the protein by incubation with 40 mM EDTA at 4 °C for at least 12 h.

Linking the Photoswitch to PDZ2. The 3,3′-bis(sulfonato)-4,4′-bis (chloroacetamido)azobenzene (BSBCA) (Fig. S1*A*) was synthesized according to the protocol from Burns et al. (1). The process to covalently link the photoswitch to two surface-exposed Cys residues, as outlined in the protocol, was modified to the following: Disulfide bridges were reduced by adding 50 mM DTT to the protein solution and incubating it for at least 1 h at room temperature. The protein was then refolded by desalting in 50 mM Tris·HCl, pH 8.5, with a HiPrep column (GE Healthcare Life Sciences), simultaneously removing the reducing agent. To prevent the reformation of disulfide bonds, fraction collection was performed in an oxygen-free (argon) atmosphere. The protein was then diluted with a well-degassed buffer (50 mM Tris·HCl, pH 8.5) to 10 μM before 50 μM of BSBCA was added, still under Ar atmosphere. The reaction vessel was sealed and stirred in the dark for 6 h. The reaction mixture was concentrated using a Vivacell pressure concentrator with a 5-kDa MWCO filter (Sartorius). To remove any surplus BSBCA the protein solution was diluted and reconcentrated twice. The monomer with a single photoswitch correctly linked to both Cys was purified using a MonoQ anion exchange column (GE Healthcare Life Sciences) in 50 mM Tris·HCl, pH 8.5, and eluted with an NaCl gradient.

His Tag Cleavage. After linking the photoswitch (BSBCA) to PDZ2, the His tag was removed with HRV 3C protease. The protease used also contained a His tag for separation from the sample. One milligram protease per 50 mg PDZ2 in 50 mM Tris·HCl, pH 8.5, was incubated at 4 °C for 16 h. The His tag-free PDZ2 was purified with a HisTrap column (GE Healthcare Life Sciences) and the buffer was exchanged by a HiPrep desalting column (GE Healthcare Life Sciences) to the corresponding buffer used for IR or NMR measurements.

Mass Spectrometry. Electrospray ionization (ESI) mass spectra were measured at the Functional Genomics Center Zurich in a mass range between 500 and 3,000 Da. The *m/z* data were deconvolved using the MasEnt1 software.

The purity of all samples was verified by SDS/PAGE (Fig. S3) and ESI mass spectrometry (Fig. S4*A*). Additionally, ESI mass spectra were repeated after time-resolved IR measurements. Thereby, we realized that the protein was modified during a measurement in Tris buffer (Fig. S4*B*). However, these modifications could be largely avoided by measuring in borate buffer and under exclusion of oxygen (Fig. S4*C*). All transient IR spectra shown in this work were therefore measured under these conditions.

Fig. S4*B* shows one modification with an increased mass of around 16 Da that can occur multiple times. We attribute this to an oxidation of the protein, most likely at the thioether groups of the cysteines linked to the photoswitch. A small part of the protein was already oxidized before the measurement (Fig. S4*A*). However, during a measurement the oxidation significantly increased. A second modification with an increased mass of 135 Da we attribute to a reaction of the oxidized protein with the Tris buffer. Both modifications increased linearly with the time of a measurement and must have been induced by the high intensity of the 420-nm laser pulses. No decomposition products with a lighter mass were detected.

IR Spectroscopy. For FTIR measurements the protein was desalted into 50 mM Tris·HCl, 150 mM NaCl, pH 8.5. We learned that in this buffer the protein is modified under the influence of 420-nm laser pluses (mass spectra, Fig. S4). Therefore, we used 50 mM borate buffer and 150 mM NaCl, pH 8.5, for time-resolved measurements. The protein was concentrated in an Amicon Ultra 3-kDa MWCO centrifugal filter device (Millipore Corporation) to 1.3 mM, lyophilized, and dissolved in D_2O. Incubation in D_2O for 2 h at room temperature before the measurements eliminated H/D exchange during experiments. FTIR spectra were measured on a Bruker Tensor 27 FTIR spectrometer in

a 50-μm or 100-μm path-length sample cell with CaF_2 windows, either in the dark or under illumination with a 150-mW, 370-nm continuous wave (cw) diode laser (CrystaLaser). The unlinked photoswitch (Figs. 1*A* and 3) was measured in the same buffer and at 1.1 mM concentration.

All FTIR and transient IR spectra were measured at room temperature (21–22 °C).

NMR Spectroscopy. For NMR spectroscopy the samples (in 50 mM sodium phosphate, 150 mM NaCl, pH 6.8) were concentrated to 0.75 mM. The buffer was exchanged (50 mM sodium phosphate, pH 6.8, 150 mM NaCl, 10% D_2O) in two successive rounds of 1:1 dilution and concentration. For measurements with the photoswitch *in cis* a 150-mW, 370-nm cw diode laser (CrystaLaser) coupled to a fiber with a custom-designed inline radial illumination probe (Polymicro Technologies) (Fig. S5) was used.

Proton chemical shifts were calibrated to the water signal and nitrogen shifts were referenced indirectly to liquid NH_3 (3). All 2D experiments used TPPI-States for quadrature detection in indirect proton dimensions and gradient-selected coherence selection (echo–antiecho) (4) in combination with sensitivity enhancement schemes in experiments including detection of amide protons.

Resonance Assignment. All NMR experiments were recorded at 25 °C on a Bruker Avance 600- or 700-MHz spectrometer equipped with a cryoprobe. Backbone resonances were identified with the help of CBCA(CO)NH (5), HNCACB (6), and HNCO (7) experiments. Side-chain assignments were obtained using HC(C)H-TOCSY (8, 9) and CC(CO)NH (10) experiments. Aromatic spin systems were linked to the backbone via a CB(CGCD)HD experiment (11). Val and Leu methyl groups were stereospecifically assigned (2). Spectra were processed under the Bruker Topspin 2.1 software and then transferred to Cara (12) for further analysis. The overall backbone/sidechain assignment completeness is 98/85% and 96/84% for the *cis* and *trans* forms, respectively. The lower percentage of backbone resonance assignments of the *trans* form is due to a bigger number of exchange-broadened peaks in proximity to the linker attachment points.

Structure Calculation. ^{15}N- and ^{13}C-edited NOESY (13) spectra with mixing times of 75 ms were used for obtaining NOE structural restraints. Two sets of amide proton residual dipolar couplings (NH-RDCs) (14) were measured for both the *cis* and the *trans* form in 8.5 mg/mL filamentous phage Pf1 (15) (ASLA Biotech) and in liquid crystalline media formed with *n*-dodecyl-penta(ethylene glycol) (C12E5) and *n*-hexanol (16). NOESY peaks were picked with ATNOS/CANDID (17, 18). Torsion angle restraints were generated with TALOS+ (19) and final structures were calculated with CYANA (20, 21). For the *trans* form a number of artificial restraints were imposed on the photoswitch to enforce a *trans*-diazo bond and coplanar aromatic rings whereas for the *cis* form only the *cis*-diazo bond was enforced (Table S2).

Structure Refinement. Structures were refined in TIP3P (three-site transferrable intermolecular potential) explicit water, using the Charmm27 force field as described previously (22–24). Parameters for the BSBCA cross-linker were derived by analogy as for azobenzene (25, 26). The structures of the *cis* and the *trans* form both show the β6/α2 architecture characteristic of PDZ2 domains. The *trans* form suffers from a number of exchange-broadened peaks, in particular in the β1-β2 loop resulting in low numbers of available distance restraints for that region concomitant with a less well-defined structure in that stretch compared with the overall structure. We point out here that the determination of the NMR structures relies on a force field similar to the one used for carrying out the molecular dynamics (MD) simulations and that distance values determined from the two methods might thus be not fully independent.

Structure Validation. Structures were validated with WHATCHECK (27, 28) and PROCHECK-NMR (29) (Table S3).

Data Bank Accession Codes. Coordinates and chemical shifts of the *cis* and *trans* configurations have been deposited in the Protein Data Bank (PDB) and the BioMagResBank (BMRB) data bank under PDB ID codes 2M0Z and 2M10 and BMRB accession nos. 18833 and 18834, respectively.

Computation. MD simulations were performed with the Gromacs program package (30) and the Gromacs implementation of the Charmm27 force field (23, 24), with a timestep of 2 fs, saving time 500 fs, all bonds constrained, isothermal–isobaric (NPT) ensemble at 300 K and 1 bar, with time constants of 0.2 ps and 0.5 ps, respectively, for the thermostat and the barostat. Lennard–Jones interactions were treated with a cutoff of 1.0 nm (switched to zero at 0.9 nm), and the long-range electrostatic forces were approximated by the Particle–Mesh–Ewald approximation. The photoswitch was parameterized as in refs. 25 and 26. To force the photoswitch to be either in the *cis* or in the *trans* configuration, the double-minimum potential for the central C-N = N-C dihedral angle was replaced by a single-minimum potential, and the force constant was increased by a factor of 3 such that the potential around the minimum agrees reasonably well with that of the original double-minimum potential from refs. 31 and 32.

PDB entry 3PDZ was used as a starting structure (33) and mutated to Ser21Cys and Glu76Cys to provide an anchor point for the photoswitch. In addition, a Tyr36Trp mutation was included in accordance with an earlier version of the experimental sequence. We inserted a photoswitch without the sulfonate groups. We were able to synthesize proteins with that version of the photoswitch in small yields and verified that despite its hydrophobicity, the protein was still folded (however, the yields of the synthesis were too small to retrieve enough material for NMR and time-resolved IR spectroscopy).

As in the experiment, we first prepared the photoswitch in its *cis* configuration. The protein was solvated in a box of 5,355 TIP3P water molecules and one Cl^- counter-ion to neutralize the simulation box, minimized with the backbone atoms constrained, and then equilibrated for 1.1 μs. From a subsequent 3-μs *cis*-equilibrium simulation, 300,000 snapshots separated by 10 ps each were taken as starting points for the nonequilibrium MD runs with the photoswitch switched into *trans*. Simulation times varied between 4 ps and 100 ns such that the number of samples in each time bin on a logarithmic time axis was roughly the same. That is, we ran 150,000 trajectories for 4 ps, 75,000 trajectories for 8 ps, and so on, up to 73 trajectories for 8 ns. In addition, 120 trajectories for a full 100 ns were collected. The total simulation time of these nonequilibrium trajectories amounts to ∼21 μs and took about 5 months on a 96-core computer cluster. Both Figs. 4 and 5 are calculated from that full set of nonequilibrium trajectories.

To calculate the water density on the protein surface (Fig. 5 and Movie S1), the 300,000 starting points, which resemble a *cis*-equilibrium ensemble, were aligned on each other by minimizing the rmsd of the C_α atoms from ordered secondary structure motives. An averaged structure was calculated from these starting points by averaging atom positions, which served as a reference to which all subsequent nonequilibrium structures were aligned together with their surrounding water molecules. The positions of the water-oxygens were then binned into cubes of 1 $Å^3$.

The residence times of water molecules were estimated by computing the average lifetime of individual water/protein hydrogen bonds along a 600-ns equilibrium MD trajectory with the photoswitch *in trans*. Time series of presence/rupture of individual hydrogen bonds were obtained by Gromacs with default thresholds of 0.35 nm for the donor–acceptor distance and 30°

for the acceptor–donor–hydrogen angle. Seven hydrogen bond donor or acceptor groups within the binding groove and 13 on the outside surface were used. Averaging over these groups resulted in hydrogen bond lifetimes of 40 ± 20 ps within the binding groove and 15 ± 5 ps on the outside surface. Neglecting transient hydrogen bond ruptures up to 10 ps yielded residence times of 60 ± 50 ps within the binding groove and 20 ± 10 ps on the outside surface.

The implicit solvent simulations were carried out according to a similar protocol with Newtonian dynamics at 300 K, using a generalized Born formalism (34). A *cis* structure already equilibrated in explicit solvent was further equilibrated for 300 ns in implicit solvent. A 700-ns production equilibrium run was then used to collect an ensemble of initial structures every 10 ns for the subsequent nonequilibrium simulations.

1. Burns DC, Zhang F, Woolley GA (2007) Synthesis of 3,3′-bis(sulfonato)-4,4′-bis (chloroacetamido)azobenzene and cysteine cross-linking for photo-control of protein conformation and activity. *Nat Protoc* 2(2):251–258.
2. Neri D, Szyperski T, Otting G, Senn H, Wüthrich K (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. *Biochemistry* 28(19):7510–7516.
3. Live DH, Davis DG, Agosta WC, Cowburn D (1984) Observation of 1000-fold enhancement of nitrogen-15 NMR via proton-detected multiquantum coherences: Studies of large peptides. *J Am Chem Soc* 106(20):6104–6105.
4. Kay LE, Keifer P, Saarinen T (1992) Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J Am Chem Soc* 114(26):10663–10665.
5. Grzesiek S, Bax A (1992) Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance nmr. *J Am Chem Soc* 114:6291–6293.
6. Grzesiek S, Bax A (1992) An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *J Magn Reson* 99:201–207.
7. Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J Magn Reson* 89:496–514.
8. Kay LE, Ikura M, Bax A (1990) Proton-proton correlation via carbon-carbon couplings: A three-dimensional NMR approach for the assignment of aliphatic resonances in proteins labeled with carbon-13. *J Am Chem Soc* 112:888–889.
9. Ikura M, Kay LE, Bax A (1991) Improved three-dimensional 1H-13C-1H correlation spectroscopy of a 13C-labeled protein using constant-time evolution. *J Biomol NMR* 1(3):299–304.
10. Grzesiek S, Anglister J, Bax A (1993) Correlation of backbone amide and aliphatic side-chain resonances in 13C/15N-enriched proteins by isotropic mixing of 13C magnetization. *J Magn Reson* 101(1):114–119.
11. Yamazaki T, Forman-Kay JD, Kay LE (1993) 2-dimensional NMR experiments for correlating C-13-beta and H-1-delta/epsilon chemical-shifts of aromatic residues in C-13-labeled proteins via scalar couplings. *J Am Chem Soc* 115:11054–11055.
12. Keller RL (2004) *The Computer Aided Resonance Tutorial* (Cantina).
13. Marion D, Kay LE, Sparks SW, Torchia DA, Bax A (1989) Three-dimensional heteronuclear NMR of nitrogen-15 labeled proteins. *J Am Chem Soc* 111(4):1515–1517.
14. Ottiger M, Delaglio F, Bax A (1998) Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J Magn Reson* 131(2):373–378.
15. Hansen MR, Hanson P, Pardi A (2000) Filamentous bacteriophage for aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. *Methods Enzymol* 317:220–240.
16. Rückert M, Otting G (2000) Alignment of biological macromolecules in novel nonionic liquid crystalline media for nmr experiments. *J Am Chem Soc* 122(32): 7793–7797.
17. Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* 319(1):209–227.
18. Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* 24(3):171–189.
19. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* 44(4):213–223.
20. Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* 273(1):283–298.
21. Güntert P (2004) Automated NMR structure calculation with CYANA. *Methods Mol Biol* 278:353–378.
22. Spronk CAEM, Linge JP, Hilbers CW, Vuister GW (2002) Improving the quality of protein structures derived by NMR spectroscopy. *J Biomol NMR* 22(3):281–289.
23. Mackerell AD, Jr., Feig M, Brooks CL, 3rd (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11):1400–1415.
24. Mackerell AD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616.
25. Spörlein S, et al. (2002) Ultrafast spectroscopy reveals subnanosecond peptide conformational dynamics and validates molecular dynamics simulation. *Proc Natl Acad Sci USA* 99(12):7998–8002.
26. Carstens H (2004) Konformationsdynamik lichtschaltbarer peptide: Molekulardynamiksimulationen und datengetriebene modellbildung [Conformational dynamics of photoswitchable peptides: Molecular dynamics and data-driven model building]. PhD thesis (Ludwig Maximilians Universität München, Munich). Available at http://edoc.ub.uni-muenchen.de/2268. German.
27. Vriend G (1990) WHAT IF: A molecular modeling and drug design program. *J Mol Graph* 8(1):52–56, 29.
28. Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. *Nature* 381(6580):272.
29. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 8(4):477–486.
30. Van Der Spoel D, et al. (2005) GROMACS: Fast, flexible, and free. *J Comput Chem* 26(16):1701–1718.
31. Nguyen PH, Stock G (2006) Nonequilibrium molecular dynamics simulation of a photoswitchable peptide. *Chem Phys* 323:36–44.
32. Ihalainen JA, et al. (2008) α-Helix folding in the presence of structural constraints. *Proc Natl Acad Sci USA* 105(28):9588–9593.
33. Kozlov G, Gehring K, Ekiel I (2000) Solution structure of the PDZ2 domain from human phosphatase hPTP1E and its interactions with C-terminal peptides from the Fas receptor. *Biochemistry* 39(10):2572–2580.
34. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55(2):383–394.

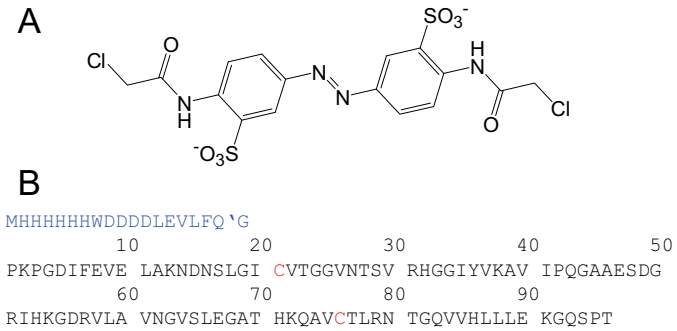


Fig. S1. Cross-linker and anchor sites in the protein sequence. (*A*) Photoswitch. This water-soluble thiol-reactive azobenzene derivative 3,3′-bis(sulfonato)-4,4′-bis(chloroacetamido)azobenzene (BSBCA) (1) was cross-linked to two cysteines in PDZ2. (*B*) Amino acid sequence of the second PDZ domain in human tyrosine-phosphatase 1E (hPTP1E). A His tag with a HRV 3C cleavage site (blue) was used for purification and cleaved before measurements. The cleavage site is marked with an apostrophe and the two Cys (S21C and E76C) that served as anchor points for the photoswitch are shown in red.

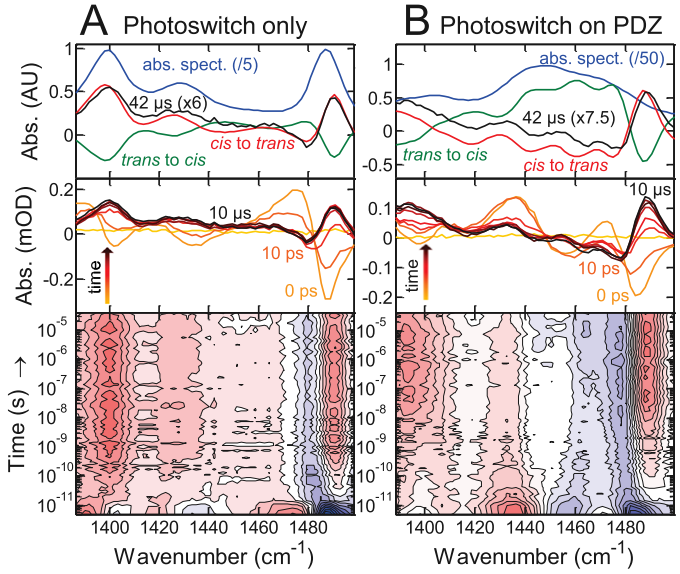


Fig. S2. IR spectra of the ring-mode (~1,390 cm^{-1}) and Amide II (~1,490 cm^{-1}) bands from the photoswitch. (*A* and *B*) (*Top*) Absolute (blue) and difference FTIR spectra (red and green) compared with the transient spectrum at 42 µs (black). (*Middle*) Transient spectra, at −1 ns (yellow), 0 s (light orange), and from 10 ps to 10 µs by decade (orange to black). (*Bottom*) Contour plot of the IR response. (*A*) Photoswitch alone; (*B*) photoswitch on PDZ2.

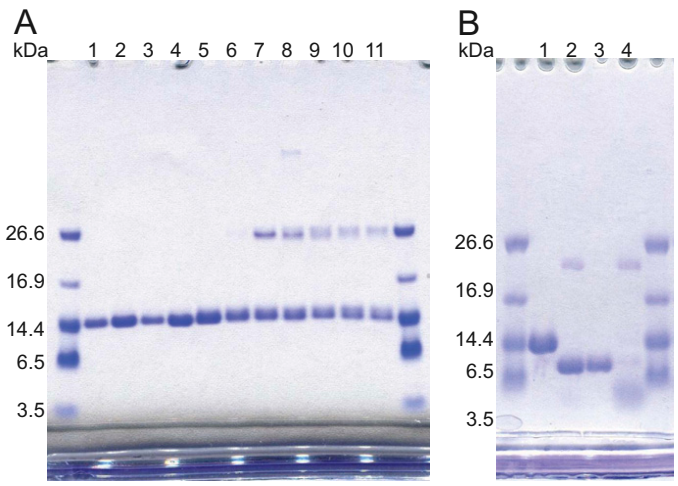


Fig. S3. SDS/PAGE analysis of linking the photoswitch (BSBCA) to PDZ2. Shown are 12% Bis-Tris gels (Novex NuPAGE; Invitrogen) with polypeptide SDS/PAGE molecular weight standards (Bio-Rad) Coomassie stained with Rotiphorese Blue R (Carl Roth GmbH). (*A*) After purification with a MonoQ anion exchange column (GE Healthcare Life Sciences). Lanes 1–5: PDZ2 with the photoswitch correctly linked to the two Cys (mass verified by ESI mass spectrometry). Lanes 6–11: Mixtures of incorrectly linked PDZ2 (monomers with two linkers, dimers, trimers). (*B*) Cleaving of the His tag. Lane 1, before cleaving; lane 2, after cleaving; lane 3, after cleaving and purification with HisTrap column (GE Healthcare Life Sciences); lane 4, HRV 3C protease and free His tag eluted from the HisTrap column.

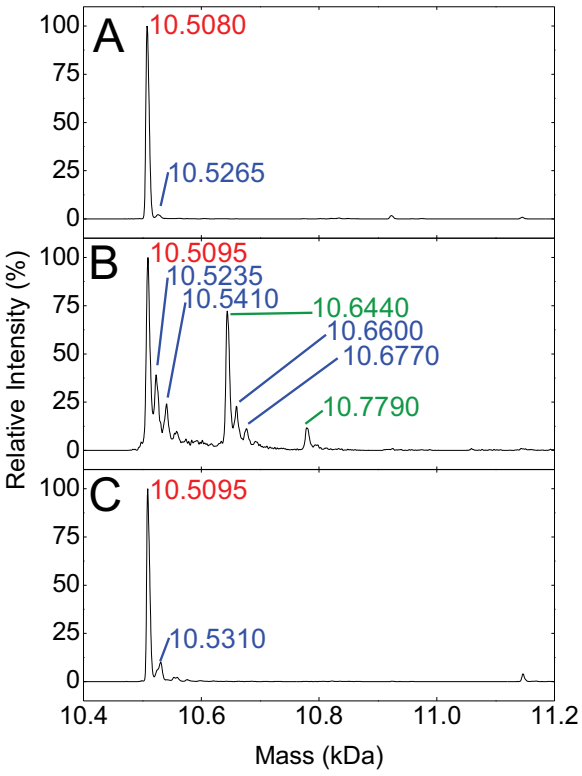


Fig. S4. Mass spectra of PDZ2 linked to the photoswitch. The peak corresponding to the expected mass is labeled in red, peaks originating from the oxidized sample are in blue, and peaks with an additional modification (+135 Da) are in green. (*A*) Before a time-resolved IR measurement, in 50 mM Tris·HCl, 150 mM NaCl, pH 8.5. (*B*) After a time-resolved IR measurement, in 50 mM Tris·HCl, 150 mM NaCl, pH 8.5. (*C*) After a time-resolved IR measurement, in 50 mM borate buffer, 150 mM NaCl, pH 8.5, and under exclusion of oxygen.

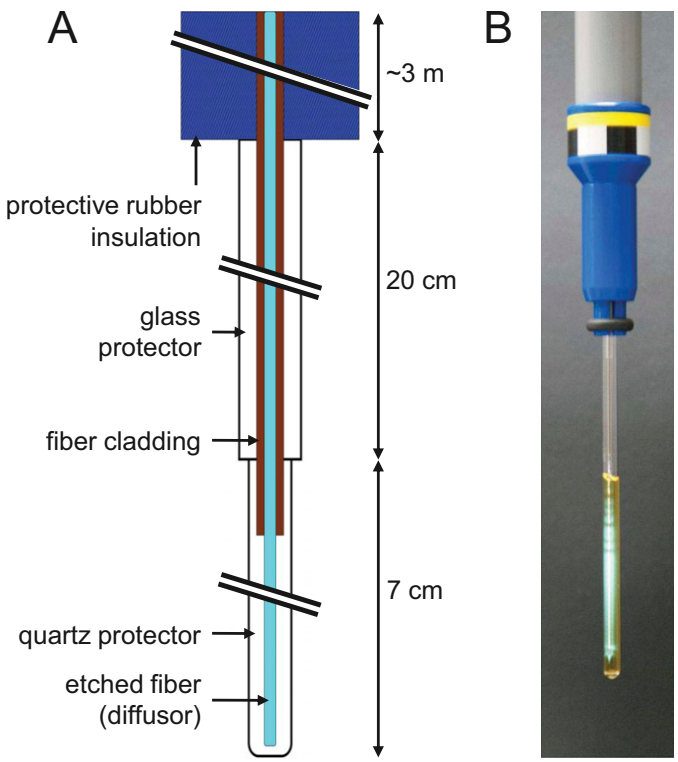


Fig. S5. Illumination of NMR sample. (*A*) Sketch of the NMR inline radial illumination probe (Polymicro Technologies). (*B*) Illumination of a sample in an NMR tube.

Table S1. Complete set of fit parameters for the black lines in Fig. 3

Probe wavenumber	a_0	a_1	τ_1	a_2	τ_2	β	a_3	τ_3
1,491 cm^{-1}	0.12	−0.16	0.015	−0.10	7.3	0.49	0	—
1,640 cm^{-1}	−0.07	0.0	0.015	0.09	7.3	0.49	−0.07	20,000

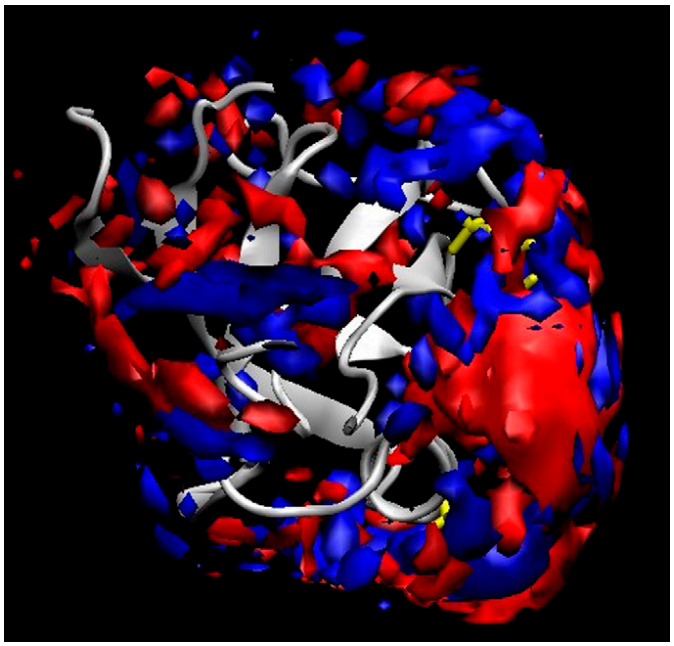
Amplitudes are in units of mOD (optical density), and time constants are in nanoseconds.

Table S2. Restraints for NMR structure calculation

Type	*cis*	*trans*
NOE	877	992
Photoswitch	2	29
Dihedral restraints	146	145
HN-RDCs		
Pf1	78	80
C12E5	76	76

Table S3. NMR structure statistics

Structural parameters of 20 NMR conformers	*cis*	*trans*
Pairwise cartesian rmsd, Å		
Global backbone heavy atoms	1.02	1.15
Global all heavy atoms	1.39	1.54
Ordered backbone heavy atoms	0.35	0.33
Ordered all heavy atoms	0.92	1.04
Ramachandran quality parameters, %		
Residues in most favored regions	86.7	86.3
Residues in allowed regions	11.2	8.8
Residues in additionally allowed regions	1.8	4.6
Residues in disallowed regions	0.3	0.3
Average rms deviation from current reliable structures, rms *Z*-scores		
Bond lengths	1.050	1.126
Bond angles	1.155	1.187
Omega angle restraints	1.303	1.486
Side-chain planarity	0.447	0.676
Improper dihedral distribution	0.908	0.983
Inside/outside distribution	0.951	0.986
Average deviation from current reliable structures, *Z*-scores		
First-generation packing quality	−1.833	−2.479
Ramachandran plot appearance	−2.014	−2.217
Chi-1/Chi-2 rotamer normality	−3.735	−3.859
Backbone conformation	−0.916	−0.916



Movie S1. Change of water density after photo-switching the azobenzene moiety that cross-links the binding groove of the PDZ2 domain. Red depicts increased water density; blue depicts decreased water density. The protein is shown by a gray ribbon; the photoswitch (visible only in part) is shown in yellow. The perturbation of the water network on the protein surface propagates within 100 ns around the protein. See Fig. 5 for details.

Movie S1

Chapter 7

Conclusions

Technological advances have accelerated the rate of data production in both computational and experimental sciences enormously. The resultant large amounts of data emerging from present-day scientific studies pose a formidable challenge in terms of data storage, data transmission and data processing. Regarding the latter, the scientific community desires scalable protocols for analysis, visualisation, and automated knowledge discovery that fully exploit the available data.

We have developed two methods for data processing to address this need in the context of computational biophysics. More specifically, we have focused on data from molecular dynamics (MD) simulations, a technique that is routinely used to investigate the complex temporal behaviour of biomolecules. The first method we have developed is a novel protocol to analyse long and high-dimensional trajectories produced by MD simulations. The method is used to generate a Sapphire (States And Pathways Projected with HIgh REsolution) plot, which illustrates the thermodynamics and the kinetics of the system in terms of its metastable states and the connectivities among them. The second method is a feature weighting scheme that can improve the performance of subsequently applied analysis algorithms.

We have thoroughly tested Sapphire plots on diverse data sets, covering model systems (diffusion on the Müller potential and n-butane) as well as peptides and a protein studied extensively in the literature (two three-stranded β-sheet peptides and bovine pancreatic trypsin inhibitor). Our tests have demonstrated both the versatility and the specific advantages of Sapphire plots, namely that the method is scalable, that distinct states do not overlap along the progress index, and that resolution is maximal. Subsequently, we have employed Sapphire plots to study peptide binding to a PDZ domain. Our analyses have shed light on the binding process and on the structure of the encounter complex that is elusive to experiments.

For Sapphire plots, there is room for future work in the following areas. First, to make the method even more efficient on modern hardware, the algorithm that generates Sapphire plots will be parallelised. Second, Sapphire plots should become more robust. The plots are based on the progress index, which is a walk through the data that proceeds through regions of high sampling density one after another. Currently, it can be beneficial

to compute the progress index several times for different starting points due to the stochasticity of the approximate algorithm and the weakness of the kinetic annotation function. Ongoing developments aim to improve the clustering-based preorganisation of the data as well as the algorithm to generate the progress index in order to make Sapphire plots more robust. As an alternative strategy, several progress indices, starting from different snapshots, could be computed, and the information could be combined in a clever way to obtain a richer aggregate picture. Furthermore, novel annotation functions might facilitate the interpretation of Sapphire plots, and protocols for finding structural features that explain basins along the progress index would prove exceptionally useful. Another avenue for future research is the automated definition of states based on one or multiple Sapphire plots.

It might be worthwhile to investigate whether Sapphire plots can be extended to visualise Markov models. A progress index of the nodes of a Markov model could be defined based on the kinetic distance among them. While the resultant plot would be similar to cut-based free energy profiles in terms of annotation functions, overlap of distinct basins would be avoided due to the different ordering of nodes. Such a representation could be very useful for the exploratory analysis of Markov models and for evaluating and comparing methods for coarse-graining Markov models.

Sapphire plots will only find widespread use if convincing examples of their contribution to the solution of biologically relevant problems are provided. Finding these problems will therefore be as important as the methodological improvements that have just been made out. We believe that the work on peptide binding to PDZ domains presented in Chapter 4 is one such example. Our study showed that electrostatic interactions play a crucial role in forming and maintaining the encounter complex. It is not clear, however, how these electrostatic interactions affect the overall binding rate. In the future, it will be feasible to investigate this question in silico by mutating charged protein residues that interact with the peptide and/or by varying the solution conditions. Sapphire plots will certainly be helpful for an efficient, yet detailed comparison of MD trajectories of such a large-scale simulation study.

We stress that Sapphire plots are not restricted to data from computational biophysics. Applying Sapphire plots to experimental data might contribute to closing the gap between simulation and experiment. The progress index is based on geometric criteria only and can thus be applied to nontemporal data in combination with application-specific annotation functions. Used in this way, Sapphire plots could play a similar role as (density-based) clustering algorithms. In contrast to most of those, Sapphire plots are scalable, do not feature a density threshold controlling what constitutes a cluster, and naturally offer a way to visualise the outcome.

Advances in computer hardware not only increase the time scales that can be sampled by MD simulations, but they also enable the simulation of ever larger systems. Simulation data are therefore constantly growing in dimensionality, and dedicated tools for data preprocessing are needed to cope with the challenges inherent to high-dimensional data. We have developed and tested a data-driven method for feature weighting to facilitate analysis of high-dimensional MD trajectories with various analysis protocols,

including SAPPHIRE plots and clustering-based approaches. Data preprocessing is of utmost importance in almost all unsupervised learning applications, and feature weighting, as proposed here, constitutes a useful strategy in general. We have proposed global and locally adaptive weights both of which are based on temporal characteristics of the data. This obviously limits their use to data recorded in time. In their current form, they can be useful in studying other complex systems exhibiting metastability. We close by pointing out that both types of weights could also be rendered locally adaptive in a geometric sense, thus considerably broadening the applicability of the approach to nontemporal data.

Acknowledgements

First and foremost, I thank Amedeo Caflisch and Andreas Vitalis. Amedeo allowed me to change my field of work from mathematics to computational biophysics and thus gave me the opportunity to broaden my scientific horizon considerably. Working under his cordial and uncomplicated leadership has always been a great pleasure. Andreas took me under his wing early on, and I benefited enormously from his guidance. He has helped me with everything from dealing with scientific everyday issues to scrutinising my own work. I wish him all the best for his scientific career. Many thanks go to the members of my thesis committee, Prof. Madhavi Krishnan and Prof. Reinhard Furrer, for their interest in my work.

I am deeply indebted to Fiona Straehl, my family, and my friends for the good times we have had, the adventures we have shared and for the support they have given me during the last years. I am grateful to all my colleagues for valuable time spent together, especially to Sandra Steiner, Andrea Magno, Riccardo Scalco, Emilie Frugier, Jean-Rémy Marchand, Dimitrios Spiliotopoulos, Lisa Gartenmann, Lisa Caflisch, Carmen Esposito, Lars Wiedmer, and Ursina Suter.

I thank the people with whom I worked and collaborated, namely Min Xu and the group of Prof. Peter Hamm, and I acknowledge David E. Shaw, Francesco Rao, and Ting Zhou for sharing data that was extremely helpful for my projects.

List of publications

Kinetic response of a photo-perturbed allosteric protein.
Buchli, B., Waldauer, S.A., Walser, R., Donten, M., Pfister, R., Blöchliger, N., Steiner, S., Caflisch, A., Zerbe, O. and Hamm, P.
[*Proceedings of the National Academy of Sciences*, 110(29): 11725–11730, 2013]

A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems.
Blöchliger, N., Vitalis, A. and Caflisch, A.
[*Computer Physics Communications*, 184(11): 2446–2453, 2013]

High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations.
Blöchliger, N., Vitalis, A. and Caflisch, A.
[*Scientific Reports*, 4: 6264, 2014]

Peptide binding to a PDZ domain by electrostatic steering via non-native salt bridges.
Blöchliger, N., Xu, M. and Caflisch, A.
[*Biophysical Journal*, 108(9): 2362–2370, 2015]

Weighted distance functions improve analysis of molecular dynamics simulation data.
Blöchliger, N., Caflisch, A. and Vitalis, A.
[*Journal of Chemical Theory and Computation*, 11(11): 5481–5492, 2015]

Curriculum Vitae

Nicolas Blöchliger

born 10th February 1987 in Aarau, Switzerland

from Unterägeri ZG

Education

September 2011 - May 2015	PhD student in the group of Prof. Dr. Amedeo Caflisch, Department of Biochemistry, University of Zurich
August 2011	Master's degree in mathematics at the University of Zurich
October 2006 - August 2011	Studies of mathematics, biology and computer science at the University of Zurich and the Université Paris Diderot in Paris, France
August 1999 - June 2005	High School Diploma at Kantonsschule Zug in Zug, Switzerland